

Approximate Bregman near neighbors in sublinear time: Beyond the triangle inequality

Amirali Abdullah
University of Utah

John Moeller
University of Utah

Suresh Venkatasubramanian
University of Utah

Abstract

Bregman divergences are important distance measures that are used extensively in data-driven applications such as computer vision, text mining, and speech processing, and are a key focus of interest in machine learning. Answering *nearest neighbor* (NN) queries under these measures is very important in these applications and has been the subject of extensive study, but is problematic because these distance measures lack metric properties like symmetry and the triangle inequality.

In this paper, we present the first provably *approximate nearest-neighbor* (ANN) algorithms. These process queries in $O(\log n)$ time for Bregman divergences in fixed dimensional spaces. We also obtain $\text{polylog } n$ bounds for a more abstract class of distance measures (containing Bregman divergences) which satisfy certain structural properties. Both of these bounds apply to both the regular asymmetric Bregman divergences as well as their symmetrized versions. To do so, we develop two geometric properties vital to our analysis: a *reverse triangle inequality* (RTI) and a relaxed triangle inequality called μ -defectiveness where μ is a domain-dependent parameter. Bregman divergences satisfy the RTI but *not* μ -defectiveness. However, we show that the square root of a Bregman divergence does satisfy μ -defectiveness. This allows us to then utilize both properties in an efficient search data structure that follows the general two-stage paradigm of a ring-tree decomposition followed by a quad tree search used in previous near-neighbor algorithms for Euclidean space and spaces of bounded doubling dimension.

Our first algorithm resolves a query for a d -dimensional $(1 + \epsilon)$ -ANN in $O\left(\left(\frac{\log n}{\epsilon}\right)^{O(d)}\right)$ time and $O(n \log^{d-1} n)$ space and holds for generic μ -defective distance measures satisfying a RTI. Our second algorithm is more specific in analysis to the Bregman divergences and uses a further structural constant, the maximum ratio of second derivatives over each dimension of our domain (c_0). This allows us to locate a $(1 + \epsilon)$ -ANN in $O(\log n)$ time and $O(n)$ space, where there is a further $(c_0)^d$ factor in the big-Oh for the query time.

1 Introduction

The nearest neighbor problem is one of the most extensively studied problems in data analysis. The past 20 years has seen tremendous research into the problem of computing near neighbors efficiently as well as approximately in different kinds of metric spaces.

An important application of the nearest-neighbor problem is in querying content databases (images, text, and audio databases, for example). In these applications, the notion of similarity is based on a distance metric that arises from information-theoretic or other considerations. Popular examples include the Kullback-Leibler divergence [16], the Itakura-Saito distance [20] and the Mahalanobis distance [27]. These distance measures are examples of a general class of divergences called the *Bregman divergences* [9], and this class has received much attention in the realm of machine learning, computer vision and other application domains.

Bregman divergences possess a rich geometric structure but are not metrics in general, and are not even symmetric in most cases! While the geometry of Bregman divergences has been studied from a combinatorial perspective and for clustering, there have been no algorithms with provable guarantees for the fundamental problem of nearest-neighbor search. This is in contrast with extensive *empirical* study of Bregman-based near-neighbor search [10, 33, 34, 36, 37].

In this paper we present the first provably approximate nearest-neighbor (ANN) algorithms for Bregman divergences. Our first algorithm processes queries in $O(\log^d n)$ time using $O(n \log^d n)$ space and only uses general properties of the underlying distance function (which includes Bregman divergences as a special case). The second algorithm processes queries in $O(\log n)$ time using $O(n)$ space and exploits structural constants associated specifically with Bregman divergences. An interesting feature of our algorithms is that they extend the “ring-tree + quad-tree” paradigm for ANN searching beyond Euclidean distances and metrics of bounded doubling dimension to distances that might not even be symmetric or satisfy a triangle inequality.

1.1 Overview of Techniques

At a high level [35], low-dimensional Euclidean approximate near-neighbor search works as follows. The algorithm builds a quad-tree-like data structure to search the space efficiently at query time. Cells reduce exponentially in size, and so a careful application of the triangle inequality and some packing bounds allows us to bound the number of cells explored in terms of the “spread” of the point set (the ratio of the maximum to minimum distance). Next, terms involving the spread are eliminated by finding an initial crude approximation to the nearest neighbor. Since the resulting depth to explore is bounded by the logarithm of the ratio of the cell sizes, any c -approximation of the nearest neighbor results in a depth of $O(\log(c/\epsilon))$. A standard data structure that yields such a crude bound is the *ring tree* [26].

Unfortunately, these methods (which work also for doubling metrics [14, 26, 7]) require two key properties: the existence of the triangle inequality, as well as packing bounds for fitting small-radius balls into large-radius balls. Bregman divergences in general are not symmetric and do not even satisfy a directed triangle inequality! We note in passing that such problems do not occur for the *exact* nearest neighbor problem in constant dimension: this problem reduces to point location in a Voronoi diagram, and Bregman Voronoi diagrams possess the same combinatorial structure as Euclidean Voronoi diagrams [8].

Reverse Triangle Inequality The first observation we make is that while Bregman divergences do not satisfy a triangle inequality, they satisfy a weak *reverse triangle inequality*: along a line, the sum of lengths of two contiguous intervals is always *less* than the length of the union. This immediately yields a packing bound: intuitively, we cannot pack too many disjoint intervals in a larger interval because their sum would then be too large, violating the reverse triangle inequality.

μ -defectiveness The second idea is to allow for a *relaxed* triangle inequality. We do so by defining a distance measure to be μ -*defective* w.r.t a given domain if there exists a fixed $\mu \geq 1$ such that for all triples of points x, y, z , we have that $|D(x, y) - D(x, z)| \leq \mu D(y, z)$. This notion was first employed by Farago et.al [19] for an algorithm based on optimizing average case complexity.

A different natural way to relax the triangle inequality would be to show there exists a fixed $\mu < 1$ such that for all triples (x, y, z) , the inequality $D(x, y) + D(y, z) \geq \mu D(x, z)$. In fact, this is the notion of μ -similarity used by Ackermann *et al* [3] to *cluster* data under a Bregman divergence. However, this version of a relaxed triangle inequality is too weak for the nearest-neighbor problem, as we see in Figure 1.

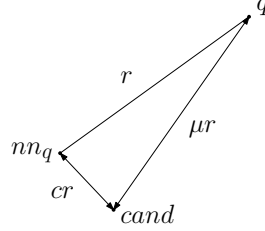


Figure 1: The ratio $\frac{D(q, \text{cand})}{D(q, \text{nn}_q)} = \mu$, no matter how small c is

Let q be a query point, cand be a point from P such that $D(q, \text{cand})$ is known and nn_q be the actual nearest neighbor to q . The principle of grid related machinery is that for $D(q, \text{nn}_q)$ and $D(q, \text{cand})$ sufficiently large, and $D(\text{cand}, \text{nn}_q)$ sufficiently small, we can verify that $D(q, \text{cand})$ is a $(1 + \varepsilon)$ nearest neighbor, i.e we can short-circuit our grid.

The figure 1 illustrates a case where this short-circuit may not be valid for μ -similarity. Note that μ -similarity is satisfied here for any $c < 1$. Yet the ANN quality of cand , i.e, $\frac{D(q, \text{cand})}{D(q, \text{nn}_q)}$, need not be better than μ even for arbitrarily close nn_q and cand ! This demonstrates the difficulty of naturally adapting the Ackermann notion of μ -similarity to finding a $1 + \varepsilon$ nearest neighbor.

In fact, the relevant relaxation of the triangle inequality that we require is slightly different. Rearranging terms, we instead require that there exist a parameter $\mu \geq 1$ such that for all triples (x, y, z) , $|D(x, y) - D(x, z)| \leq \mu D(y, z)$. We call such a distance μ -defective. It is fairly straightforward to see that a μ -defective distance measure is also $2/(\mu + 1)$ -similar, but the converse does not hold, as the example above shows.

Without loss of generality, assume that $D(x, y) \geq D(x, z) \geq D(y, z)$. Then $D(x, y) - D(x, z) \leq \mu D(y, z)$ and $D(x, y) - D(y, z) \leq \mu D(x, z)$, so $2D(x, y) \leq (\mu + 1)(D(x, z) + D(y, z))$. Since $D(x, y)$ is the greatest of the three distances, this inequality is the strongest and implies the corresponding $2/(\mu + 1)$ -similarity inequalities for the other two distances.

Unfortunately, Bregman divergences do not satisfy μ -defectiveness for any size domain or value of μ ! One of our technical contributions is demonstrating in Section 4 that surprisingly, the square root of Bregman divergences does satisfy this property with μ depending on the boundedness of the domain and choice of divergence.

A Generic Approximate Near-Neighbor Algorithm After establishing that Bregman divergences satisfy the reverse triangle inequality and μ -defectiveness (Section 4), we first show (Section 6) that *any* distance measure satisfying the reverse triangle inequality, μ -defectiveness, and some mild technical conditions admits a ring-tree-based construction to obtain a weak near neighbor. However, applying it to a quad-tree construction creates a problem. The μ -defectiveness of a distance measure means that if we take a unit length interval and divide it into two parts, all we can expect is that each part has length between $1/2$ and $1/(\mu + 1)$. This implies that while we may have to go down to level $\lceil \log_2 \ell \rceil$ to guarantee that all cells have side length $O(\ell)$, some cells might have side length as little as $\ell^{\log_2(\mu+1)}$, weakening packing bounds considerably.

We deal with this problem in two ways. For Bregman divergences, we can exploit geometric properties of the associated convex function ϕ (see Section 3) to ensure that cells at a fixed level have bounded size (Section 8); this is achieved by reexamining the second derivative ϕ'' .

For more general abstract distances that satisfy the reverse triangle inequality and μ -defectiveness, we

instead construct a portion of the quad tree “on the fly” for each query (Section 7). While this is expensive, it still yields $\text{polylog}(n)$ bounds for the overall query time in fixed dimensions. Both of these algorithms rely on packing/covering bounds that we prove in Section 5.

An important technical point is that for exposition and simplicity, we initially work with the *symmetrized* Bregman divergences (of the form $D_{s\phi}(x, y) = D_\phi(x | y) + D_\phi(y | x)$), and then extend these results to general Bregman divergences (Section 9). We note that the results for symmetrized Bregman divergences might be interesting in their own right, as they have also been used in applications [33, 34, 32, 30].

2 Related Work

Approximate nearest-neighbor algorithms come in two flavors: the high dimensional variety, where all bounds must be polynomial in the dimension d , and the constant-dimensional variety, where terms exponential in the dimension are permitted, but query times must be sublinear in n . In this paper, we focus on the constant-dimensional setting. The idea of using ring-trees appears in many works [25, 26, 23], and a good exposition of the general method can be found in Har-Peled’s textbook [35, Chapter 11].

The Bregman distances were first introduced by Bregman[9]. They are the unique divergences that satisfy certain axiom systems for distance measures [17], and are key players in the theory of information geometry [5]. Bregman distances are used extensively in machine learning, where they have been used to unify boosting with different loss functions[15] and unify different mixture-model density estimation problems [6]. A first study of the algorithmic geometry of Bregman divergences was performed by Nielsen, Nock and Boissonnat [8]. This was followed by a series of papers analyzing the behavior of clustering algorithms under Bregman divergences [3, 2, 1, 28, 13].

Many heuristics have also been proposed for spaces endowed with Bregman divergences. Nielsen and Nock [31] developed a Frank-Wolfe-like iterative scheme for finding minimum enclosing balls under Bregman divergences. Cayton [10] proposed the first nearest-neighbor search strategy for Bregman divergences, based on a clever primal-dual branch and bound strategy. Zhang *et al* [37] developed another prune-and-search strategy that they argue is more scalable and uses operations better suited to use within a standard database system. For good broad reviews of near neighbor search in theory and practice, the reader is referred to the books by Har-Peled[35], Samet [24] and Shakhnarovich *et al* [29].

3 Definitions

In this paper we study the approximate nearest neighbor problem for distance functions D : Given a point set P , a query point q , and an error parameter ϵ , find a point $\text{nn}_q \in P$ such that $D(\text{nn}_q, q) \leq (1 + \epsilon) \min_{p \in P} D(p, q)$. We start by defining general properties that we will require of our distance measures. In what follows, we will assume that the distance measure D is *reflexive*: $D(x, y) = 0$ iff $x = y$.

Definition 3.1 (Monotonicity). *Let $M \subset \mathbb{R}$, $D : M \times M \rightarrow \mathbb{R}$ be a distance function, and let $a, b, c \in M$ where $a < b < c$. If the following are true for any such choice of a, b , and c : that $0 \leq D(a, b) < D(a, c)$, that $0 \leq D(b, c) < D(a, c)$, and that $D(x, y) = 0$ iff $x = y$, then we say that D is monotonic.*

For a general distance function $D : M \times M \rightarrow \mathbb{R}$, where $M \subset \mathbb{R}^d$, we say that D is monotonic if it is monotonic when restricted to any subset of M parallel to a coordinate axis.

Definition 3.2 (Reverse Triangle Inequality). *Let M be a subset of \mathbb{R} . We say that a monotone distance measure $D : M \times M \rightarrow \mathbb{R}$ satisfies a reverse triangle inequality or RTI if for any three elements $a \leq b \leq c \in M$, $D(a, b) + D(b, c) \leq D(a, c)$*

Definition 3.3 (μ -defectiveness). *Let D be a symmetric monotone distance measure satisfying the reverse triangle inequality. We say that D is μ -defective with respect to domain M if for all $a, b, q \in M$,*

$$|D(a, q) - D(b, q)| < \mu D(a, b) \quad (3.1)$$

For an asymmetric distance measure D , we define left and right sided μ -defectiveness respectively as

$$|D(q, a) - D(q, b)| < \mu D(a, b) \quad (3.2)$$

$$|D(a, q) - D(b, q)| < \mu D(b, a) \quad (3.3)$$

Note that by interchanging a and b and using the symmetry of the modulus sign, we can also rewrite left and right sided μ -defectiveness respectively as $|D(q, a) - D(q, b)| < \mu D(b, a)$ and $|D(a, q) - D(b, q)| < \mu D(a, b)$.

Two technical notes. The distance functions under consideration are typically defined over \mathbb{R}^d . We will assume in this paper that the distance D is *decomposable*: roughly, that $D((x_1, \dots, x_d), (y_1, \dots, y_d))$ can be written as $g(\sum_i f(x_i, y_i))$, where g and f are monotone. This captures all the Bregman divergences that are typically used (with the exception of the Mahalanobis distance). We will also need to compute the diameter of an axis parallel box of side length ℓ . Our results hold as long as the diameter of such a box is $O(\ell d^{O(1)})$: note that this captures standard distances like those induced by norms, as well as decomposable Bregman divergences. In what follows, we will mostly make use of the *square root* of a Bregman divergence, for which the diameter of a box is $\ell(\mu + 1)d^{\frac{1}{2}}$ or $\ell d^{\frac{1}{2}}$, and so without loss of generality we will use this in our bounds.

Bregman Divergences. Let $\phi : M \subset \mathbb{R}^d \rightarrow \mathbb{R}$ be a *strictly convex* function that is differentiable in the relative interior of M . The *Bregman divergence* D_ϕ is defined as

$$D_\phi(x, y) = \phi(x) - \phi(y) - \langle \nabla \phi(y), x - y \rangle \quad (3.4)$$

In general, D_ϕ is asymmetric. A *symmetrized* Bregman divergence can be defined by averaging:

$$D_{s\phi}(x, y) = \frac{1}{2}(D_\phi(x, y) + D_\phi(y, x)) = \frac{1}{2}\langle x - y, \nabla \phi(x) - \nabla \phi(y) \rangle \quad (3.5)$$

An important subclass of Bregman divergences are the *decomposable* Bregman divergences. Suppose ϕ has domain $M = \prod_{i=1}^d M_i$ and can be written as $\phi(x) = \sum_{i=1}^d \phi_i(x_i)$, where $\phi_i : M_i \subset \mathbb{R} \rightarrow \mathbb{R}$ is also strictly convex and differentiable in $\text{relint}(S_i)$. Then $D_\phi(x, y) = \sum_{i=1}^d D_{\phi_i}(x_i, y_i)$ is a *decomposable* Bregman divergence.

Most commonly used Bregman divergences are decomposable: [11, Chapter 3] illustrates some of the commonly used ones, including the Euclidean distance, the KL-divergence, and the Itakura-Saito distance. In this paper we will hence limit ourselves to considering decomposable distance measures. We note that due to the primal-dual relationship of $D_\phi(a, b)$ and $D_{\phi^*}(b^*, a^*)$, for our results on the asymmetric Bregman divergence we need only consider right-sided μ -defective distance measures.

4 Properties of Bregman Divergences

The previous section defined key properties that we desire of a distance function D . The Bregman divergences (or modifications thereof) satisfy the following properties, as can be shown by direct computation.

Lemma 4.1. *Any one-dimensional Bregman divergence is monotonic.*

Lemma 4.2. *Any one-dimensional Bregman divergence satisfies the reverse triangle inequality. Let $a \leq b \leq c$ be three points in the domain of D_ϕ . Then it holds that:*

$$D_\phi(a, b) + D_\phi(b, c) \leq D_\phi(a, c) \quad (4.1)$$

$$D_\phi(c, b) + D_\phi(b, a) \leq D_\phi(c, a) \quad (4.2)$$

This is also true for $D_{s\phi}$ and $\sqrt{D_{s\phi}}$.

Note that this lemma can be extended similarly by induction to any series of n points between a and c . Further, using the relationship between $D_\phi(a, b)$ and the “dual” distance $D_{\phi^*}(b^*, a^*)$, we can show that the reverse triangle inequality holds going “left” as well: $D_\phi(c, b) + D_\phi(b, a) \leq D_\phi(c, a)$. These two separate reverse triangle inequalities together yield the result for $D_{s\phi}$. The corresponding proof for $\sqrt{D_{s\phi}}$ merely requires some algebraic manipulation.

For the remainder of this section, the proofs are straightforward but tedious, and hence we consign them to Appendix A. While the Bregman divergences satisfy both monotonicity and the reverse triangle inequality, they are not μ -defective with respect to *any* domain! An easy example of this is ℓ_2^2 , which is also a Bregman divergence. A surprising fact however is that $\sqrt{D_{s\phi}}$ and $\sqrt{D_\phi}$ do satisfy μ -defectiveness (with μ depending on the bounded size of our domain). While we were unable to show precise bounds for μ in terms of the domain, the values are small. For example, for the symmetrized KL-divergence on the simplex where each coordinate is bounded between 0.1 and 0.9, μ is 1.22. If each coordinate is between 0.01 and 0.99, then μ is 2.42.

Lemma 4.3. *Given any interval $I = [x_1, x_2]$ on the real line, there exists a finite μ such that $\sqrt{D_{s\phi}}$ is μ -defective with respect to I , and $\sqrt{D_\phi}$ is both left and right-sided μ -defective with respect to I .*

We note that the result for $\sqrt{D_\phi}$ is proven by establishing the following relationship between $D_\phi(a, b)$ and $D_\phi(b, a)$ over a bounded interval $I \subset \mathbb{R}$, and with some further computation.

Lemma 4.4. *Given a Bregman divergence D_ϕ and a bounded interval $I \subset \mathbb{R}$, $\sqrt{D_\phi(a, b)}/\sqrt{D_\phi(b, a)}$ is bounded by a constant $c_0 \forall a, b \in I$ where c_0 depends on the choice of divergence and interval.*

We extend our results to d dimensions naturally now by showing that if M is a domain such that $\sqrt{D_{s\phi}}$ and $\sqrt{D_\phi}$ are μ -defective with respect to the projection of M onto each coordinate axis, then $\sqrt{D_{s\phi}}$ and $\sqrt{D_\phi}$ are μ -defective with respect to all of M .

5 Packing and Covering Bounds

The aforementioned key properties (monotonicity, the reverse triangle inequality, decomposability, and μ -defectiveness) can be used to prove packing and covering bounds for a distance measure D . We now present some of these bounds.

Lemma 5.1 (Interval packing). *Consider a monotone distance measure D satisfying the reverse triangle inequality, an interval $[ab]$ such that $D(a, b) = s$ and a collection of disjoint intervals intersecting $[ab]$, where $I = \{[xx'] \mid [xx'], D(x, x') \geq \ell\}$. Then $|I| \leq \frac{s}{\ell} + 2$.*

Proof. Let I' be the intervals of I that are totally contained in $[ab]$. The combined length of all intervals in I' is at most $|I'|\ell$, but by the reverse triangle inequality, their total length cannot exceed s , so $|I'| \leq \frac{s}{\ell}$. There can be only two members of I not in I' , so $|I| \leq \frac{s}{\ell} + 2$. \square

A simple greedy approach yields a constructive version of this lemma.

Corollary 5.1. *Given any two points, $a \leq b$ on the line s.t $D(a, b) = s$, we can construct a packing of $[ab]$ by $r \leq \frac{1}{\varepsilon}$ intervals $[x_i, x_{i+1}]$, $1 \leq i \leq r$ such that $D(a, x_0) = D(x_i, x_{i+1}) = \varepsilon s$, $\forall i$ and $D(x_r, b) \leq \varepsilon s$. Here D is a monotone distance measure satisfying the reverse triangle inequality.*

Recall here that D_ϕ , $D_{s\phi}$ and $\sqrt{D_{s\phi}}$ satisfy the conditions of Lemma 5.1 and corollary 5.1 as they satisfy an RTI and are decomposable. However, since $\sqrt{D_\phi}$ may not satisfy the reverse triangle inequality, we instead prove a weaker packing bound on $\sqrt{D_\phi}$ by using D_ϕ .

Lemma 5.2 (Weak interval packing). *Given distance measure $\sqrt{D_\phi}$ and an interval $[ab]$ such that $\sqrt{D_\phi}(a, b) = s$ and a collection of disjoint intervals intersecting $[ab]$ where $I = \{[xx'] \mid [xx'], \sqrt{D_\phi}(x, x') \geq \ell\}$. Then $|I| \leq \frac{s^2}{\ell^2} + 2$. Such a set of intervals can be explicitly constructed.*

Proof. We note that here $D_\phi(a, b) = s^2$, and $I = \{[xx'] \mid [xx'], D_\phi(x, x') \geq \ell^2\}$. The result then follows trivially from lemma 5.1, since D_ϕ satisfies the conditions of lemma 5.1. \square

The above bounds can be generalized to higher dimensions to provide packing bounds for balls and cubes (which we define below) with respect to a monotone, decomposable distance measure.

Definition 5.1. Given a collection of d intervals a_i, b_i , s.t $D(a_i, b_i) = s$ where $1 \leq i \leq d$, the cube in d dimensions is defined as $\prod_{i=1}^d [a_i b_i]$ and is said to have side length s .

Lemma 5.3. Given a d dimensional cube B_1 of side length s under distance measures D_ϕ , $D_{s\phi}$ and $\sqrt{D_{s\phi}}$, we can cover it with at most ϵ^d cubes of side length exactly ϵs . In the case of $\sqrt{D_\phi}$, we can cover it with at most ϵ^{2d} cubes of side length ϵs .

Proof. Note that $D_{s\phi}$, D_ϕ , $\sqrt{D_{s\phi}}$ satisfy conditions of corollary 5.1. Hence we can construct a gridding of at most $\frac{1}{\epsilon}$ points in each dimension spaced ϵs apart. We then take a product over all d dimensions, and the lemma follows trivially. For $\sqrt{D_\phi}$, we refer to the RTI for and follow the same procedure, gridding by at most $\frac{1}{\epsilon^2}$ points in each dimension, spaced ϵs apart. \square

Lemma 5.4. Consider a ball B of radius s and center C with respect to a distance measure D . Then in the case of $D_{s\phi}$ and D_ϕ it can be covered with $\frac{2^d}{\epsilon^d}$ balls of radius $d\epsilon s$. In the case of $\sqrt{D_{s\phi}}$, B can be covered with $\frac{2^d}{\epsilon^d}$ balls of radius $\sqrt{d}\epsilon s$. And for $\sqrt{D_\phi}$, B can be covered by $\frac{2^d}{\epsilon^{2d}}$ balls of radius $\sqrt{d}\epsilon s$.

Proof. We divide the ball into 2^d orthants around the center c . Each orthant can be covered by a cube of size s . We now consider each case separately. For $D_{s\phi}$, D_ϕ and $\sqrt{D_{s\phi}}$, by lemma 5.3 each such cube can be broken down into $\frac{1}{\epsilon^d}$ sub-cubes of side length ϵs . For $\sqrt{D_\phi}$, we can break down each cube into $\frac{1}{\epsilon^{2d}}$ sub-cubes of side length ϵs .

For $D_{s\phi}$ we can trivially cover each sub-cube by a ball of radius $d\epsilon s$ placed at any corner. Similarly, for $\sqrt{D_{s\phi}}$, we can cover each sub-cube by a ball of radius $\sqrt{d}\epsilon s$ placed at any corner. (This latter result follows by considering the sub-cube of side length ϵs under $\sqrt{D_{s\phi}}$ as one of side length $\epsilon^2 s^2$ under $D_{s\phi}$ and placing a ball of radius $d\epsilon^2 s^2$ under $D_{s\phi}$ on any corner).

We now consider the cases of D_ϕ and $\sqrt{D_\phi}$. For each orthant, we construct the gridding by Lemmas 5.1 and 5.2 in each dimension for D_ϕ and $\sqrt{D_\phi}$ respectively. This gives us d sets of points X_i , $1 \leq i \leq d$, where X_i lies on the i -th axis passing through the center of the ball C . For each X_i , we have an ordering (by construction) of points C, x_{i1}, x_{i2}, \dots , s.t $D(x_i, x_{i+1}) = \epsilon s$. Clearly every subcube is induced by the product of d pairs of points of the form $\{x_{i(m_i-1)}, x_{im_i}\}$ where $1 \leq i \leq d$ and m_i is some positive integer. Now to each subcube assign the lowest corner L_c , defined as the product of the points $x_{i(m_i-1)}$, $1 \leq i \leq d$. The Bregman ball of radius $d\epsilon s$ with center L_c will cover this subcube for the case D_ϕ , and the Bregman ball of radius $\sqrt{d}\epsilon s$ with center L_c will cover this subcube for the case $\sqrt{D_\phi}$. Note that this argument will also extend to covering the cells of a quadtree produced by recursive decomposition, by a ball of required size placed on appropriate "lowest" corner.

Since there are $\frac{1}{\epsilon^{2d}}$ and $\frac{1}{\epsilon^d}$ sub-cubes to each orthant for $\sqrt{D_\phi}$ and D_ϕ respectively, the lemma now follows by covering each subcube with a Bregman ball of the required radius. \square

6 Computing a rough approximation

Armed with our packing and covering bounds, we now describe how to compute a $O(\log n)$ rough approximate nearest-neighbor on our point set P , which we will use in the next section to find the $(1 + \epsilon)$ -approximate nearest neighbor. The technique we use is based on ring separators. Ring separators are a fairly old concept in geometry, notable appearances of which include the landmark paper by Indyk and Motwani [25]. Our

approach here is heavily influenced by Har-Peled and Mendel [23], and by Krauthgamer and Lee [26], and our presentation is along the template of the textbook by Har-Peled [35, Chapter 11].

We note here that the constant of $d^{d/2}$ which appears in our final bounds for storage and query time is specific to $\sqrt{D_{s\phi}}$. However, an argument on the same lines will yield a constant of $d^{O(d)}$ for any generic μ -defective, symmetric RTI-satisfying decomposable distance measure such that the diameter of a cube of side length 1 is bounded by $d^{O(1)}$.

Let $B(m, r)$ denote the ball of radius r centered at m , and let $B'(m, r)$ denote the complement (or exterior) of $B(m, r)$. A *ring* R is the difference of two concentric balls: $R = B(m, r_2) \setminus B(m, r_1), r_2 \geq r_1$. We will often refer to the larger ball $B(m, r_2)$ as B_{out} and the smaller ball as B_{in} . We use $P_{\text{out}}(R)$ to denote the set $P \cap B_{\text{out}}$, and similarly use $P_{\text{in}}(R)$ as $P \cap B_{\text{in}}$, where we may drop the reference to R when the context is obvious. A *t-ring separator* $R_{P,c}$ on a point set P is a ring such that $\frac{n}{c} < |P_{\text{in}}| < (1 - \frac{1}{c})n, \frac{n}{c} < |P_{\text{out}}| < (1 - \frac{1}{c})n, r_2 \geq (1+t)r_1$ and $B_{\text{out}} \setminus B_{\text{in}}$ is empty. A *t-ring tree* is a binary tree obtained by repeated dispartition of our point set P using a *t-ring separator*.

Note that later on in this section, we will abuse this notation slightly by using ring-separators where the annulus is not actually empty, but we will bound the added space complexity and tree depth introduced

Finally, denote the minimum sized ball containing at least $\frac{n}{c}$ points of P by $B_{\text{opt},c}$; its radius is denoted by $r_{\text{opt},c}$.

We demonstrate that for any point set P , a ring separator exists and secondly, it can always be computed efficiently. Applying this “separator” recursively on our point structure yields a ring-tree structure for searching our point set. Before we proceed further, we need to establish some properties of disks under a μ -defective distance. Lemma 6.1 is immediate from the definition of μ -defectiveness, Lemma 6.2 is similar to one obtained by Har-Peled and Mazumdar [22] and the idea of repeating points in both children of a ring-separator derives from a result by Har-Peled and Mendel [23].

Lemma 6.1. *Let D be a μ -defective distance, and let $B(m, r)$ be a ball with respect to D . Then for any two points $x, y \in B(m, r)$, $D(x, y) < (\mu + 1)r$.*

Lemma 6.2. *Given a constant $1 \leq c \leq n$, we can compute in $O(nc)$ randomized time a $\mu + 1$ approximation to the smallest radius ball containing $\frac{n}{c}$ points.*

Proof. As described by Har Peled-Mazumdar ([22]) we let S be a random sample from P , generated by choosing every point of P with probability $\frac{c}{n}$. Next, compute for every $p \in S$, the smallest disk centered at p containing c points. By median selection, this can be done in $O(n)$ time and since $E(|S|) = c$, this gives us the expected running time of $O(nc)$. Now, let r' be the minimum radius computed. Note that by lemma 6.1, if $|S \cap B_{\text{opt},c}| > 0$ then we have that $r' \leq (\mu + 1)r_{\text{opt}}$. But since $B_{\text{opt},c}$ contains $\frac{n}{c}$ points, we can upper bound the probability of failure as the probability that we do not select any of the $\frac{n}{c}$ points in B_{opt} in our sample. Hence:

$$\Pr(|S \cap B_{\text{opt},c}| > 0) = 1 - (1 - \frac{c}{n})^{\frac{n}{c}} \geq 1 - \frac{1}{e}$$

Note that one can obtain a similar approximation deterministically by brute force search, but this would incur a prohibitive $O(n^2)$ running time. \square

We can now use Lemma 6.2 to construct our ring-separator.

Lemma 6.3. *For arbitrary t s.t $1 < t < n$, we can construct a $\frac{1}{t}$ -ring separator $R_{P,c}$ in $O(n)$ expected time on a point set P by repeating points.*

Proof. Using Lemma 6.2, we compute a ball $S = B(m, r_1)$ (where $m \in P$) containing $\frac{n}{c}$ points such that $r_1 \leq (\mu + 1)r_{\text{opt},c}$ where c is a constant to be set. Consider the ball $\bar{S} = B(m, 2r_1)$. We shall argue that

there must be $\frac{n}{c}$ points of P in \bar{S}' , for careful choices of c . As described in Lemma 5.4, \bar{S} can be covered by 2^d hypercubes of side length $2r_1$, the union of which we shall refer to as H . Set $L = (\mu + 1)\sqrt{d}$. Imagine a partition of H into a grid, where each cell is of side-length $\frac{r_1}{L}$ and hence of diameter at most $\Delta(\frac{r_1}{L}, d) = \frac{r_1}{\mu+1} \leq r_{\text{opt},c}$. A ball of radius $r_{\text{opt},c}$ on any corner of a cell will contain the entire cell, and so it will contain at most $\frac{n}{c}$ points, by the definition of $r_{\text{opt},c}$.

By Lemma 5.3 the grid on H has at most $2^d(2r_1/\frac{r_1}{L})^d = (4(\mu + 1)\sqrt{d})^d$ cells. Set $c = 2(4(\mu + 1)\sqrt{d})^d$. Then we have that $\bar{S} \subset H$ contains at most $\frac{n}{c}(4(\mu + 1)\sqrt{d})^d = \frac{n}{2}$ points. Since the inner ball S contains at least $\frac{n}{c}$ points, and the outer ball \bar{S} contains at most $\frac{n}{2}$ points, hence the annulus $\bar{S} \setminus S$ contains at most $\frac{n}{2} - \frac{n}{c}$ points. Now, divide $\bar{S} \setminus S$ into t rings of equal width, and by the pigeonhole principle at least one of these rings must contain at most $O(\frac{n}{t})$ points of P . Now let the inner ball corresponding to this ring be B_{in} and the outer ball be B_{out} and add these points to *both* children. Even for $t = 1$, each child contains at most $\frac{n}{2} + (\frac{n}{2} - \frac{n}{c}) = (1 - \frac{1}{c})n$ points. Also, the thickness of the ring is bounded by $\frac{2r_1 - r_1}{t}/2r_1 = \frac{1}{2t}$, i.e it is a $O(\frac{1}{t})$ ring separator. Finally, we can check in $O(n)$ time if the randomized process of Lemma 6.2 succeeded simply by verifying the number of points in the inner and outer ring. \square

Lemma 6.4. *Given any point set P , we can construct a $O(\frac{1}{\log n})$ ring-separator tree T of depth $O(d^{\frac{d}{2}}(\mu + 1)^d \log n)$.*

Proof. Repeatedly partition P by lemma 6.2 into P_{in}^v and P_{out}^v where v is the parent node. Store only the single point $\text{rep}_v = m \in P$ in node v , the center of the ball $B(m, r_1)$. We continue this partitioning until we have nodes with only a single point contained in them. Since each child contains at least $\frac{n}{c}$ points (by proof of Lemma 6.3), each subset reduces by a factor of at least $1 - \frac{1}{c}$ at each step, and hence the depth of the tree is logarithmic. We calculate the depth more exactly, noting that in Lemma 6.3, $c = O(d^{\frac{d}{2}}(\mu + 1)^d)$. Hence the depth x can be bounded as:

$$\begin{aligned} n(1 - \frac{1}{c})^x &= 1 \\ (1 - \frac{1}{c})^x &= \frac{1}{n} \\ x &= \frac{\ln \frac{1}{n}}{\ln(1 - \frac{1}{c})} = \frac{-1}{\ln(1 - \frac{1}{c})} \ln n \\ x &\leq c \ln n = O\left(d^{\frac{d}{2}}(\mu + 1)^d \log n\right) \end{aligned}$$

\square

Finally, we verify that the storage space require is not excessive.

Lemma 6.5. *To construct a $O(\frac{1}{\log n})$ ring-separator tree requires $O(n)$ storage and $O(d^{\frac{d}{2}}(\mu + 1)^d n \log n)$ time.*

Proof. By Lemma 6.4 the depth bounds still hold upon repeating points. For storage, we have to bound the total number of points in our data structure after repetition, let us say P_R . Since each node corresponds to a splitting of P_R , there may be only $O(P_R)$ nodes and total storage. Note in the proof of Lemma 6.3, for a node containing x points, at most an additional $\frac{x}{\log n}$ may be duplicated in the two children.

To bound this over each level of our tree, we sum across each node to obtain that the number of points T_i at the i -th level, as:

$$T_i = T_{i-1} \left(1 + \frac{1}{\log T_{i-1}}\right) \quad (6.1)$$

Note also by Lemma 6.4, the tree depth is $O(\log n)$ or bounded by $k \log n$ where k is a constant. Hence we only need to bound the storage at the level $i = O(\log n)$. We solve the recurrence, noting that $T_0 = n$ and $T_i > n$ for all i and hence $T_i < T_{i-1}(1 + \frac{1}{\log n})$. Thus the recurrence works out to:

$$T_i < n \left(1 + \frac{1}{\log n}\right)^{O(\log n)} < n \left(\left(1 + \frac{1}{\log n}\right)^{\log n}\right)^k < n(e^k).$$

Where the main algebraic step is that $(1 + \frac{1}{x})^x < e$. This proves that the number of points, and hence our storage complexity is $O(n)$. Multiplying the depth by $O(n)$ for computing the smallest ball across nodes on each level, gives us the time complexity of $O(n \log n)$. We note that other tradeoffs are available for different values of approximation quality (t) and construction time / query time. \square

Algorithm and Quality Analysis Let best_q be the best candidate for nearest neighbor to q found so far and $D_{\text{near}} = D(\text{best}_q, q)$. Let nn_q be the exact nearest neighbor to q from point set P and $D_{\text{exact}} = D(\text{nn}_q, q)$ be the exact nearest neighbor distance. Finally, let **curr** be the tree node currently being examined by our algorithm, and rep_{curr} be a representative point $p \in P$ of **curr**. By convention r_v represents the radius of the *inner* ball associated with a node v , and within each node v we store $\text{rep}_v = m_v$, which is the center of $B_{\text{in}}(m_v, r_v)$. The node associated with the inner ball B_{in} is denoted by v_{in} and the node associated with B_{out} is denoted by v_{out} .

Lemma 6.6. *Given a t -ring tree T for a point set with respect to a μ -defective distance D , we can find a $O(\mu + \frac{2\mu^2}{t})$ nearest neighbor in $O((\mu + 1)^d d^{\frac{d}{2}} \log n)$ time.*

Proof. Our search algorithm is a binary tree search. Whenever we reach node v , if $D(\text{rep}_v, q) < D_{\text{near}}$ set $\text{best}_q = \text{rep}_v$ and $D_{\text{near}} = D(\text{rep}_v, q)$ as our current nearest neighbor and nearest neighbor distance respectively. Our branching criterion is that if $D(\text{rep}_v, q) < (1 + \frac{t}{2})r_v$, we continue search in v_{in} , else we continue the search in v_{out} . Since the depth of the tree is $O(\log n)$ by Lemma 6.4, this process will take $O(\log n)$ time.

Turning now to quality, let w be the first node such that $\text{nn}_q \in w_{\text{in}}$ but we searched in w_{out} , or vice-versa. After examining rep_w , $D_{\text{near}} \leq D(\text{rep}_w, q)$ and D_{near} can only decrease at each step. An upper bound on $D(q, \text{rep}_w)/D(q, \text{nn}_q)$ yields a bound on the quality of the approximate nearest neighbor produced. In the first case, suppose $\text{nn}_q \in w_{\text{in}}$, but we searched in w_{out} . Then $D(\text{rep}_w, q) > (1 + \frac{t}{2})r_w$ and $D(\text{rep}_w, \text{nn}_q) < r_w$. Now μ -defectiveness implies that $\mu D(q, \text{nn}_q) > D(\text{rep}_w, q) - D(\text{rep}_w, \text{nn}_q)$, so we have $D(q, \text{nn}_q) > \frac{t}{2\mu}r_w$. And for the upper bound on $D(\text{rep}_w, q)/D(q, \text{nn}_q)$, we again apply μ -defectiveness to conclude that $D(\text{rep}_w, q) - D(q, \text{nn}_q) < \mu D(\text{nn}_q, \text{rep}_w)$, which yields $\frac{D(\text{rep}_w, q)}{D(q, \text{nn}_q)} < 1 + \mu \frac{r_w}{D(q, \text{nn}_q)} < 1 + \mu \frac{r_w}{\frac{t}{2\mu}r_w} = 1 + 2\frac{\mu^2}{t}$.

We now consider the other case. Suppose $\text{nn}_q \in w_{\text{out}}$ and we search in w_{in} instead. By construction we must have $D(\text{rep}_w, q) < (1 + \frac{t}{2})r_w$ and $D(\text{rep}_w, \text{nn}_q) > (1 + t)r_w$. Again, μ -defectiveness yields $D(q, \text{nn}_q) > \frac{t}{2\mu}r_w$. Now we can simply take the ratios of the two: $\frac{D(\text{rep}_w, q)}{D(q, \text{nn}_q)} < \frac{(1 + \frac{t}{2})r_w}{\frac{t}{2\mu}r_w} = \mu + \frac{2\mu}{t}$. Taking an upper bound of the approximation provided by each case, the ring tree provides us a $\mu + 2\frac{\mu^2}{t}$ approximation. \square

Corollary 6.1. *Setting $t = \frac{1}{\log n}$, we can find a $O(\mu + 2\mu^2 \log n)$ approximate nearest neighbor to a query point q in $O(d^{\frac{d}{2}}(\mu + 1)^d \log(n))$ time, using a $O(\frac{1}{\log n})$ ring separator tree.*

Proof. By Lemma 6.4, Lemma 6.3 and Lemma 6.6. Note that we are slightly abusing notation in Lemma 6.3, in that the separating ring obtained there is not empty of points of P as originally stipulated. However remember that if nn_q is in the ring, then nn_q repeats in *both* children and cannot fall off the search path. Hence we can “pretend” the ring is empty as in our analysis in Lemma 6.6. \square

7 Overall algorithm

We give now our overall algorithm for obtaining a $1 + \varepsilon$ nearest neighbor in $O(\frac{1}{\varepsilon^d} \log^{2d} n)$ query time.

7.1 Preprocessing

We first construct an improved ring-tree R on our point set P in $O(n \log n)$ time as described in Lemma 6.5, with ring thickness $O(\frac{1}{\log n})$. We then compute an efficient orthogonal range reporting data structure on P in $O(n \log^{d-1} n)$ time, such as that described in [4] by Afshani *et al.* We note the main result we need:

Lemma 7.1. *We can compute a data structure from P with $O(n \log^{d-1} n)$ storage (and same construction time), such that given an arbitrary axis parallel box B we can determine in $O(\log^d n)$ query time a point $p \in P \cap B$ if $|P \cap B| > 0$*

7.2 Query handling

Given a query point q , we use R to obtain a point q_{rough} in $O(\log n)$ time such that $D_{\text{rough}} = D(q, q_{\text{rough}}) \leq (1 + \mu^2 \log n) D(q, \text{nn}_q)$. Given q_{rough} , we can use Lemma 5.4 to find a family F of 2^d cubes of side length exactly D_{rough} such that they cover $B(q, D_{\text{rough}})$. We use our range reporting structure to find a point $p \in P$ for all non-empty cubes in F in a total of $2^d \log^d n$ time. These points act as representatives of the cubes for what follows. Note that nn_q must necessarily be in one of these cubes, and hence there must be a $(1 + \varepsilon)$ -nearest neighbor $q_{\text{approx}} \in P$ in some $G \in F$. To locate this q_{approx} , we construct a quadtree [35, Chapter 11] [18] for repeated bisection and search on each $G \in F$.

Algorithm 1 describes the overall procedure. We call the collection of all cells produced during the procedure a *quadtree*. We borrow the presentation in Har-Peled's book [35] with the important qualifier that we construct our quadtree at runtime. The terminology here is as introduced earlier in section 6.

Algorithm 1 QueryApproxNN(P, root, q)

```

Instantiate a queue  $Q$  containing all cells from  $F$  along with their representatives and enqueue root  $\log n$ 
Let  $D_{\text{near}} = D(\text{rep}_{\text{root}}, q)$ ,  $\text{best}_q = \text{rep}_{\text{root}}$ 
repeat
  Pull off the head of the queue and place it in curr.
  if  $D(\text{rep}_{\text{curr}}, q) < D(\text{best}_q, q)$  then
    Let  $\text{best}_q = \text{rep}_{\text{curr}}$ ,  $D_{\text{near}} = D(\text{best}_q, q)$ 
    Bisect curr according to procedure of Lemma 7.3; place the result in  $\{G_i\}$ .
    for all  $G_i$  do
      As described in 7.3, check if  $G_i$  is non-empty by passing it to our range reporting structure, which
      will also return us some  $p \in P$  if  $G_i$  is not empty.
      Also check if  $G_i$  may contain a point closer than  $(1 - \frac{\varepsilon}{2})D_{\text{near}}$  to  $q$ . (This may be done in  $O(d)$  time
      for each cell, given the coordinates of the corners.)
      if  $G_i$  is non-empty AND has a close enough point to  $q$  then
        Let  $\text{rep}_{G_i} = p$ 
        Enqueue  $G_i$ 
      end if
    end for
  end if
until  $Q$  is empty
Return  $\text{best}_q$ 

```

Lemma 7.2. *Algorithm 1 will always return a $(1 + \varepsilon)$ -approximate nearest neighbor.*

Proof. Let best_q be the point returned by the algorithm at the end of execution. By the method of the algorithm, for all points p for which the distance is directly evaluated, we have that $D(\text{best}_q, q) < D(p, q)$.

The terminology here is as in section 6. We look at points p which are *not* evaluated during the running of the algorithm, i.e. we did not expand their containing cells. But by the criterion of the algorithm for not expanding a cell, it must be that $D(\text{best}_q, q)(1 - \frac{\epsilon}{2}) < D(p, q)$. For $\epsilon < 1$, this means that $(1 + \epsilon)D(p, q) > D(\text{best}_q, q)$ for any $p \in P$, including nn_q . So best_q is indeed a $1 + \epsilon$ approximate nearest neighbor. \square

We must analyze the time complexity of a single iteration of our algorithm, namely the complexity of a subdivision of a cube G and determining which of the 2^d subcells of G are non-empty.

Lemma 7.3. *Let G be a cube with maximum side length s and G_i its subcells produced by bisecting along each side of G . For all non-empty subcubes G_i of G , we can find $p_i \in P \cap G_i$ in $O(2^d \log^d n)$ total time complexity, and the maximum side length of any G_i is at most $\frac{s}{2}$.*

Proof. Note that G is defined as a product of d intervals. For each interval, we can find an approximate bisecting point in $O(1)$ time and by the RTI each subinterval is of length at most $\frac{s}{2}$. This leads to an $O(d)$ cost to find a bisection point for all intervals, which define $O(2^d)$ subcubes or children.

We pass each subcube of G to our range reporting structure. By lemma 7.1, this takes $O(\log^d n)$ time to check emptiness or return a point $p_i \in P$ contained in the child, if non-empty. Since there are $O(2^d)$ non-empty children of G , this implies a cost of $2^d (\log^d n)$ time incurred.

Checking each of the non-empty subcubes G_i to see if it may contain a point closer than $(1 - \frac{\epsilon}{2})D_{\text{near}}$ to q takes a further $O(d)$ time per cell or $O(d2^d)$ time. \square

We now bound the number of cells that will be added to our search queue. We do so indirectly, by placing a lower bound on the maximum side length of all such cells.

Lemma 7.4. *Algorithm 1 will not add the children of node \mathbf{C} to our search queue if the maximum side length of \mathbf{C} is less than $\frac{\epsilon D(q, \text{nn}_q)}{2\mu\sqrt{d}}$.*

Proof. Let $\Delta(\mathbf{C})$ represent the diameter of cell \mathbf{C} . By construction, we can expand \mathbf{C} only if some subcell of \mathbf{C} contains a point p such that $D(p, q) \leq (1 - \frac{\epsilon}{2})D_{\text{near}}$. Note that since \mathbf{C} is examined, we have $D_{\text{near}} \leq D(\text{rep}_{\mathbf{C}}, q)$. Now assuming we expand \mathbf{C} , then we must have:

$$\mu\Delta(\mathbf{C}) > D(\text{rep}_{\mathbf{C}}, q) - D(p, q) \geq D_{\text{near}} - (1 - \frac{\epsilon}{2})D_{\text{near}} = \frac{\epsilon}{2}D_{\text{near}} \quad (7.1)$$

So $\epsilon/(2\mu)D_{\text{near}} < \Delta(\mathbf{C})$. First note $D(\text{rep}_{\mathbf{C}}, q) < D_{\text{near}}$. Also, by definition, $D(q, \text{nn}_q) < D_{\text{near}}$. And $\Delta(\mathbf{C}) < \sqrt{d}s$ where s is the maximum side length of \mathbf{C} . Making the appropriate substitutions yields us our required bound. \square

Given the bound on quadtree depth (Lemma 7.4), and using the fact that at most 2^{xd} nodes are expanded at level x , we have:

Lemma 7.5. *Given a cube G of side length D_{rough} , we can compute a $(1 + \epsilon)$ -nearest neighbor to q in $O\left(\frac{1}{\epsilon^d} 2^d \mu^d d^{\frac{d}{2}} \left(\frac{D_{\text{rough}}}{D(q, \text{nn}_q)}\right)^d \log^d n\right)$ time.*

Proof. Consider a quadtree search from q on a cube G of side length D_{rough} . By lemma 7.4, our algorithm will not expand cells with all side lengths smaller than $\frac{\epsilon D(q, \text{nn}_q)}{2\mu\sqrt{d}}$. But since the side length reduces by at least half in each dimension upon each split, all side lengths are less than this value after $x = \log\left(D_{\text{rough}} / \frac{\epsilon D(q, \text{nn}_q)}{2\mu\sqrt{d}}\right)$ repeated bisections of our root cube.

Noting that $O(\log^d n)$ time is spent at each node by lemma 7.3, and that at the x -th level the number of nodes expanded is 2^{xd} , we get a final time complexity bound of $O\left(\frac{1}{\epsilon^d} 2^d \mu^d d^{\frac{d}{2}} \left(\frac{D_{\text{rough}}}{D(q, \text{nn}_q)}\right)^d \log^d n\right)$. \square

Substituting $D_{\text{rough}} = \mu^2 \log n D(q, \text{nn}_q)$ in Lemma 7.5 gives us a bound of $O\left(2^d \frac{1}{\varepsilon^d} \mu^{3d} d^{\frac{d}{2}} \log^{2d} n\right)$. This time is per cube of F that covers $B(q, D_{\text{rough}})$. Noting that there are 2^d such cubes gives us a final time complexity of $O\left(2^{2d} \frac{1}{\varepsilon^d} \mu^{3d} d^{\frac{d}{2}} \log^{2d} n\right)$. For the space complexity of our run-time queue, observe that the number of nodes in our queue increases only if a node has more than one non-empty child, i.e., there is a split of our n points. Since our point set may only split n times, this gives us a bound of $O(n)$ on the space complexity of our queue.

8 Logarithmic bounds, with further assumptions.

For a given $D_{s\phi}$, let $c_0 = \max_{i \in [1..d]} \sqrt{\frac{\max_x \phi_i''(x)}{\min_y \phi_i''(y)}}$. We conjecture that $c_0 = \Theta(\mu)$ although we cannot prove it. In particular, we show that if we assume a bounded c_0 (in addition to μ), we can obtain a $1 + \varepsilon$ nearest neighbor in time $O(\log n + (\frac{1}{\varepsilon})^d)$ time for $\sqrt{D_{s\phi}}$. We do so by constructing a *Euclidean* quadtree T on our set in preprocessing and using c_0 and μ to express the bounds obtained in terms of $\sqrt{D_{s\phi}}$.

We will refer to the Euclidean distance l_2 as D_e and note first the following key relation between $\sqrt{D_{s\phi}}$ and D_e .

Lemma 8.1. *Suppose we are given a interval $I = [x_1 x_2] \subset \mathbb{R}$ s.t. $x_1 < x_2$, $D_e(x_1, x_2) = r_e$, and $\sqrt{D_{s\phi}(x_1, x_2)} = r_\phi$. Suppose we divide I into m subintervals of equal length with endpoints $x_1 = a_0, a_1, \dots, a_{m-1}, a_m = x_2$, where $a_i < a_{i+1}$ and $D_e(a_i, a_{i+1}) = r_e/m$, $\forall i \in [0..m-1]$. Then $\frac{r_\phi}{c_0 m} \leq \sqrt{D_{s\phi}(a_i, a_{i+1})} \leq \frac{c_0 r_\phi}{m}$.*

Proof Sketch: We can relate $\sqrt{D_{s\phi}}$ to D_e via the Taylor expansion of $\sqrt{D_{s\phi}}$: $\sqrt{D_{s\phi}(a, b)} = \sqrt{\phi''(\bar{x})} D_e(a, b)$ for some $\bar{x} \in [ab]$. Combining this with c_0 yields $\frac{\min_i \sqrt{D_{s\phi}(a_i, a_{i+1})}}{\sqrt{D_{s\phi}(x_1, x_2)}} \geq \frac{D_e(a_i, a_{i+1})}{c_0 D_e(x_1, x_2)} = \frac{1}{c_0 m}$ and $\frac{\max_i \sqrt{D_{s\phi}(a_i, a_{i+1})}}{\sqrt{D_{s\phi}(x_1, x_2)}} \leq c_0 \frac{D_e(a_i, a_{i+1})}{D_e(x_1, x_2)} = \frac{c_0}{m}$. \square

Corollary 8.1. *If we recursively bisect an interval $I = [x_1 x_2] \subset \mathbb{R}$ s.t. $D_e(x_1, x_2) = r_e$ and $\sqrt{D_{s\phi}(x_1, x_2)} = r_\phi$ into 2^i equal subintervals (under D_e), then $\frac{r_\phi}{c_0 2^i} \leq \sqrt{D_{s\phi}(a_k, a_{k+1})} \leq \frac{c_0 r_\phi}{2^i}$ for any of the subintervals $[a_k a_{k+1}]$ so obtained. Hence after $\log \frac{c_0 r_\phi}{x}$ subdivisions, all intervals will be of length at most x under $\sqrt{D_{s\phi}}$. Also, given a cube of initial side length r_ϕ , after $\log \frac{c_0 r_\phi}{x}$ repeated bisections (under D_e) the diameter will be at most $\sqrt{d}x$ under $\sqrt{D_{s\phi}}$.*

We find the smallest enclosing Bregman cube of side length s that bounds our point set, and then construct our compressed Euclidean quadtree in preprocessing. Corollary 8.1 gives us that for cells formed at the i -th level of decomposition, the side length under $\sqrt{D_{s\phi}}$ is between $\frac{s}{c_0 2^i}$ and $\frac{c_0 s}{2^i}$. Refer to these cells formed at the i -th level as L_i .

Lemma 8.2. *Given a ball B of radius r under $\sqrt{D_{s\phi}}$, let $i = \log \frac{s}{c_0 r}$. Then $|L_i \cap B| \leq O(2^d)$ and the side length of each cell in L_i is between r and $c_0^2 r$ under $\sqrt{D_{s\phi}}$. We can also explicitly retrieve the quadtree cells corresponding to $|L_i \cap B|$ in $O(2^d \log n)$ time.*

Proof. Note that for cells in L_i , we have side lengths under $\sqrt{D_{s\phi}}$ between $\frac{s}{c_0 2^i}$ and $\frac{c_0 s}{2^i}$ by Corollary 8.1. Substituting $i = \log \frac{s}{c_0 r}$, these cells have side length between r and $c_0^2 r$ under $\sqrt{D_{s\phi}}$. By the reverse triangle inequality and a similar argument to Lemma 5.4, we get our required bound for $|L_i \cap B|$. In preconstruction of our quadtree T we maintain for each dimension the corresponding interval quadtree T_k , $\forall k \in [1..d]$. Observe this incurs at most $O(n)$ storage, with d in the big-Oh. For retrieving the actual cells $|L_i \cap B|$, we first find the $O(1)$ intervals from level i in each T_k that may intersect B . Taking a product of these, we get $O(2^d)$ cells which are a superset of the canonical cells $L_i \subset T$. Each cell may be looked up in $O(\log n)$ time from the compressed quadtree [35] so our overall retrieval time is $O(2^d \log n)$. \square

Given query point q , we first obtain in $O(\log n)$ time with our ring-tree a rough $O(n)$ ANN q_{rough} s.t. $D_{\text{rough}} = \sqrt{D_{s\phi}(q, q_{\text{rough}})} = \mu^2 n \sqrt{D_{s\phi}(q, \text{nn}_q)}$. By Lemma 8.2, we have $O(2^d)$ quadtree cells intersecting $B(q, \sqrt{D_{s\phi}(q, q_{\text{rough}})})$.

Let us call this collection of cells Q . We then carry out a quadtree search on each element of Q . Note that we expand only cells which may contain a point nearer to query point q than the current best candidate. We bound the depth of our search using μ -defectiveness similar to Lemma 7.4:

Lemma 8.3. *We will not expand cells of diameter less than $\frac{\varepsilon \sqrt{D_{s\phi}(q, \text{nn}_q)}}{2\mu}$ or cells whose side-lengths w.r.t. $\sqrt{D_{s\phi}}$ are less than $\frac{\varepsilon \sqrt{D_{s\phi}(q, \text{nn}_q)}}{2\mu \sqrt{d}}$.*

For what follows, refer to our *spread* as $\beta = \frac{D_{\text{rough}}}{\sqrt{D_{s\phi}(q, \text{nn}_q)}}$.

Lemma 8.4. *We will only expand our tree to a depth of $k = \log(2c_0^3 \mu \beta \sqrt{d}/\varepsilon)$.*

Proof. Using Lemma 8.3 and Corollary 8.1, each cell of Q will be expanded only to a depth of $k = \log\left(c_0 c_0^2 D_{\text{rough}} / \frac{\varepsilon \sqrt{D_{s\phi}(q, \text{nn}_q)}}{2\mu \sqrt{d}}\right)$. This gives us a depth of $\log(2c_0^3 \mu \beta \sqrt{d}/\varepsilon)$. \square

Lemma 8.5. *The number of cells examined at the i -th level is $n_i < 2^d \left(\mu^d d^{\frac{d}{2}} c_0^{4d} + \left(\frac{2^i c_0}{\beta}\right)^d \right)$.*

Proof. Recalling that the cells of Q start with side length at most $c_0^2 D_{\text{rough}}$, at the i -th level the diameter of cells is at most $\frac{c_0^3 \sqrt{d} D_{\text{rough}}}{2^i}$ under $\sqrt{D_{s\phi}}$, by Corollary 8.1. Hence by μ -defectiveness, there must be some point examined by our algorithm at distance at most $D_{\text{best}} = \sqrt{D_{s\phi}(q, \text{nn}_q)} + \frac{\mu c_0^3 \sqrt{d} D_{\text{rough}}}{2^i}$. Note that our algorithm will only expand cells within this distance of q .

The side-length of a cell \mathbf{C} at this level is at least $\Delta(\mathbf{C}) = \frac{D_{\text{rough}}}{c_0 2^i}$. Applying the packing bounds from Lemma 5.3, and the fact that $(a+b)^d < 2^d (a^d + b^d)$, the number of cells expanded is at most

$$n_i = \left(\frac{D_{\text{best}}}{\Delta(\mathbf{C})} \right)^d < 2^d \left(\mu^d d^{\frac{d}{2}} c_0^{4d} + \left(\frac{c_0 2^i}{\beta} \right)^d \right). \quad \square$$

Finally we add the n_i to get the total number of nodes explored:

$$\sum_i n_i = O\left(2^d \mu^d d^{\frac{d}{2}} c_0^{4d} \log(2c_0^3 \mu \beta \sqrt{d}/\varepsilon) + 2^{2d} c_0^{4d} \mu^d d^{\frac{d}{2}} / \varepsilon^d\right).$$

Recalling that $\beta = \frac{D_{\text{rough}}}{\sqrt{D_{s\phi}(q, \text{nn}_q)}} = \mu^2 n$, substituting back and ignoring lower order terms, the time complexity is

$$O\left(2^d \mu^d d^{\frac{d}{2}} c_0^{4d} \log n + 2^{2d} c_0^{4d} \mu^d d^{\frac{d}{2}} / \varepsilon^d\right).$$

Accounting for the 2^d cells in Q that we need to search, this adds a further 2^d multiplicative factor. We note that compressed Euclidean quadtrees can be built in $O(n \log n)$ time and require $O(n)$ space [35], which matches our bound for the ring-tree search phase of our algorithm requiring $O(n \log n)$ time and $O(n)$ space.

9 The General Case: Asymmetric Divergences

Without loss of generality we will focus on the *right-sided* nearest neighbor: given a point set P , query point q and $\varepsilon \geq 0$, find $x \in P$ that approximates $\min_{p \in P} D(p, q)$ to within a factor of $(1 + \varepsilon)$. Since a Bregman divergence is not in general μ -defective, we will consider instead $\sqrt{D_\phi}$: by monotonicity and with an appropriate choice of ε , the result will carry over to D_ϕ .

We list three issues that have to be resolved to complete the algorithm. Firstly, because of asymmetry, we cannot bound the diameter of a quadtree cell \mathbf{C} of side length s by $s\sqrt{d}$. However, as the proof of Lemma 5.4 shows, we can choose a *canonical corner* of a cell such that a (directed) ball of radius $s\sqrt{d}$ centered at that corner covers the cell. By μ -defectiveness, we can now conclude that the diameter of \mathbf{C} is at most $(\mu + 1)s\sqrt{d}$ (note that this incurs an extra factor of $\mu + 1$ in all expressions). Secondly, since while $\sqrt{D_\phi}$ satisfies μ -defectiveness (unlike D_ϕ) the opposite is true for the reverse triangle inequality, which is satisfied by D_ϕ but not $\sqrt{D_\phi}$. This requires the use of a weaker packing bound based on Lemma 5.2, introducing dependence in $1/\varepsilon^2$ instead of $1/\varepsilon$. And thirdly, the lack of symmetry means we have to be careful of the use of directionality when proving our bounds.

Note that for this section, when we speak of asymmetric μ -defective distance measure D , we are referring to $\sqrt{D_\phi}$. With some small adjustments, similar bounds can be obtained for more generic asymmetric, monotone, decomposable and μ -defective distance measures satisfying packing bounds. The left-sided asymmetric nearest neighbor can be determined analogously.

Finally, given a bounded domain D , we have that $\sqrt{D_\phi}$ is left-sided μ_L -defective for some μ_L and right-sided μ_R -defective for some μ_R (see Lemma A.4 for detailed proof). For what follows, let $\mu = \max(\mu_L, \mu_R)$ and describe D as simply μ -defective.

Most of the proofs here mirror their counterparts in Sections 6 and 7.

9.1 Asymmetric ring-trees

Since we focus on *right-near-neighbors*, all balls and ring separators referred to will use *left-balls* i.e balls $B(m, r) = \{x \mid D(m, x) < r\}$. As in Section 6, we will design a ring-separator algorithm and use that to build a ring-separator tree.

Lemma 9.1. *Let D be a μ -defective distance, and let $B(m, r)$ be a left-ball with respect to D . Then for any two points $x, y \in B(m, r)$, $D(x, y) < (\mu + 1)r$.*

As in Lemma 6.2 we can construct (in $O(nc)$ expected time) a $(\mu + 1)$ -approximate left-ball enclosing $\frac{n}{c}$ points. This in turn yields a ring-separator construction, and from it a ring tree with an extra $(\mu + 1)^d d^{\frac{d}{2}}$ factor in depth as compared to symmetric ring-trees, due to the weaker packing bounds for $\sqrt{D_\phi}$.

We note that the asymptotic bounds for ring-tree storage and construction time follow from purely combinatorial arguments and hence are unchanged for $\sqrt{D_\phi}$. Once we have the ring-tree, we can use it as before to identify a rough near-neighbor for a query q ; once again, exploiting μ -defectiveness gives us the desired approximation guarantee for the result.

Lemma 9.2. *Given any constant $1 \leq c \leq n$, we can compute in $O(nc)$ randomized time a left-ball $B(m, r')$ such that $r' \leq (\mu + 1)r_{\text{opt}, c}$ and $B(m, r') \cap P \geq \frac{n}{c}$.*

Proof. The proof is similar enough to Lemma 6.2 that we omit details here. □

Lemma 9.3. *There exists a constant c (which depends only on d and μ), such that for any d -dimensional point set P and any μ -defective distance D , we can find a $O(\frac{1}{\log n})$ left-ring separator $R_{P, c}$.*

Proof. First, using our randomized construction, we compute a ball $S = B(m, r_1)$ (where $m \in P$) containing $\frac{n}{c}$ points such that $r_1 \leq (\mu + 1)r_{\text{opt}, c}$, where c is a constant to be set. Consider the ball $\tilde{S} = B(m, 2r_1)$. As described in Lemma 5.4, \tilde{S} can be covered by 2^d hypercubes of side length $2r_1$, the union of which we shall

refer to as H . Set $L = (\mu + 1)\sqrt{d}$. Imagine a partition of H into a grid, where each cell is of side-length $\frac{r_1}{L}$. Each cell in this grid can be covered by a ball of radius $\Delta(\frac{r_1}{L}, d) = \frac{r_1}{\mu+1} \leq r_{\text{opt},c}$ centered on it's lowest corner. This implies any cell will contain at most $\frac{n}{c}$ points, by the definition of $r_{\text{opt},c}$.

By Lemma 5.3 the grid on H has at most $2^d(2r_1/\frac{r_1}{L})^{2d} = (4(\mu + 1)\sqrt{d})^{2d}$ cells. Each cell may contain at most $\frac{n}{c}$ points. In particular, set $c = 2(4(\mu + 1)\sqrt{d})^{2d}$. Then we have that H may contain at most $\frac{n}{c}(4(\mu + 1)\sqrt{d})^{2d} = \frac{n}{2}$ points, or since $\bar{S} \subset H$, \bar{S} contains at most $\frac{n}{2}$ points and \bar{S}' contains at least $\frac{n}{2}$ points. The rest of the proof goes through as in Lemma 6.3 \square

We proceed now to the construction of our ring-tree using the basic ring-separator structure of Lemma 9.3.

Lemma 9.4. *Given any point set P , we can construct a $O(\frac{1}{\log n})$ left ring-separator tree T of depth $O(d^d(\mu + 1)^{2d} \log n)$.*

Proof. Repeatedly partition P by Lemma 9.3 into P_{in}^v and P_{out}^v where v is the parent node. Store only the single point $\text{rep}_v = m \in P$ in node v , the center of the ball $B(m, r_1)$. We continue this partitioning until we have nodes with only a single point contained in them.

Since each child contains at least $\frac{n}{c}$ points, each subset reduces by a factor of at least $1 - \frac{1}{c}$ at each step, and hence the depth of the tree is logarithmic. We calculate the depth more exactly, noting that in Lemma 9.3, $c = O(d^d(\mu + 1)^{2d})$. Hence the depth x can be bounded as:

$$\begin{aligned} n(1 - \frac{1}{c})^x &= 1 \\ (1 - \frac{1}{c})^x &= \frac{1}{n} \\ x &= \frac{\ln \frac{1}{n}}{\ln(1 - \frac{1}{c})} = \frac{-1}{\ln(1 - \frac{1}{c})} \ln n \\ x &\leq c \ln n = O(d^d(\mu + 1)^{2d} \log n) \end{aligned}$$

\square

Note that Lemma 9.4 also serves to bound the query time of our data structure. We need only now bound the approximation quality. The derivation is similar to Lemma 6.6, but with some care about directionality.

Lemma 9.5. *Given a t -ring tree T for a point set with respect to a right-sided μ -defective distance D , we can find a $O(\mu + \frac{2\mu^2}{t})$ nearest neighbor $O((\mu + 1)^d d^{\frac{d}{2}} \log n)$ time.*

Proof. Our search algorithm is a binary tree search. Whenever we reach node v , if $D(\text{rep}_v, q) < D_{\text{near}}$ set $\text{best}_q = \text{rep}_v$ and $D_{\text{near}} = D(\text{rep}_v, q)$ as our current nearest neighbor and nearest neighbor distance respectively. Our branching criterion is that if $D(\text{rep}_v, q) < (1 + \frac{t}{2})r_v$, we continue search in v_{in} , else we continue the search in v_{out} . Since the depth of the tree is $O(\log n)$ by Lemma 9.4, this process will take $O(\log n)$ time.

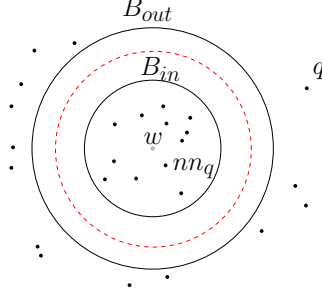


Figure 2: q is outside $(1 + \frac{t}{2})r_{in}$ so we search \mathbf{w}_{out} , but $nn_q \in \mathbf{w}_{in}$

Let \mathbf{w} be the first node such that $nn_q \in \mathbf{w}_{in}$ but we searched in \mathbf{w}_{out} , or vice-versa. The analysis goes by cases. In the first case as seen in figure 2, suppose $nn_q \in \mathbf{w}_{in}$, but we searched in \mathbf{w}_{out} . Then

$$D(\text{rep}_w, q) > \left(1 + \frac{t}{2}\right) r_w$$

$$D(\text{rep}_w, nn_q) < r_w.$$

Now left-sided μ -defectiveness implies a lower bound on the value of $D(nn_q, q)$:

$$\mu D(nn_q, q) > D(\text{rep}_w, q) - D(\text{rep}_w, nn_q)$$

$$\mu D(nn_q, q) > \left(1 + \frac{t}{2}\right) r_w - r_w$$

$$D(nn_q, q) > \frac{t}{2\mu} r_w,$$

And for the upper bound on $D(\text{rep}_w, q)/D(nn_q, q)$. First by right-sided μ -defectiveness:

$$D(\text{rep}_w, q) - D(nn_q, q) < \mu D(\text{rep}_w, nn_q)$$

$$D(\text{rep}_w, q) < D(nn_q, q) + \mu r_w$$

$$\frac{D(\text{rep}_w, q)}{D(nn_q, q)} < 1 + \mu \frac{r_w}{D(nn_q, q)}$$

$$\frac{D(\text{rep}_w, q)}{D(nn_q, q)} < 1 + \mu \frac{r_w}{\frac{t}{2\mu} r_w}$$

$$\frac{D(\text{rep}_w, q)}{D(nn_q, q)} < 1 + \mu \frac{2\mu}{t}$$

$$\frac{D(\text{rep}_w, q)}{D(nn_q, q)} < 1 + 2\frac{\mu^2}{t}$$

We now consider the other case. Suppose $nn_q \in \mathbf{w}_{out}$ and we search in \mathbf{w}_{in} instead. The analysis is almost identical. By construction we must have:

$$D(\text{rep}_w, q) < \left(1 + \frac{t}{2}\right) r_w$$

$$D(\text{rep}_w, nn_q) > (1 + t)r_w$$

Again, left-sided μ -defectiveness yields:

$$D(nn_q, q) > \frac{t}{2\mu} r_w$$

We can simply take the ratios of the two:

$$\frac{D(\text{rep}_w, q)}{D(\text{nn}_q, q)} < \frac{(1 + \frac{t}{2})r_w}{\frac{t}{2\mu}r_w} = \mu + \frac{2\mu}{t}$$

Taking an upper bound of the approximation quality provided by each case, we get that the ring separator provides us a $\mu + 2\frac{\mu^2}{t}$ rough approximation. \square

9.2 Asymmetric quadtree decomposition

As in Section 7, we use the approximate near-neighbor returned by the ring-separator-tree query to progressively expand cells, using a subdivide-and-search procedure similar to Algorithm 1. A key difference is the procedure used to bisect a cell.

Lemma 9.6. *Let G be a cube with maximum side length s and G_i its subcells produced by partitioning each side of G into two equal intervals. For all non-empty subcubes G_i of G , we can find $p_i \in P \cap G_i$ in $O(2^d \log^d n)$ total time complexity, and the maximum side length of any G_i is at most $\frac{s}{\sqrt{2}}$.*

Proof. Note that G is defined as a product of d intervals. For each interval, we can find an approximate bisecting point in $O(1)$ time. Here the bisection point x of interval $[ab]$ is such that $\sqrt{D_\phi(a, x)} = \sqrt{D_\phi(x, b)}$. By resorting to the RTI for D_ϕ , we get that $D_\phi(a, x) + D_\phi(x, b) < s^2$ and hence $D_\phi(a, x) = D_\phi(x, b) < \frac{s^2}{2}$ which implies $\sqrt{D_\phi(a, x)} = \sqrt{D_\phi(x, b)} < \frac{s}{\sqrt{2}}$. The rest of our proof follows as in Lemma 7.3 \square

We now bound the number of cells that will be added to our search queue. We do so indirectly, by placing a lower bound on the maximum side length of all such cells, and note that for the asymmetric case we get an additional factor of $\frac{1}{\mu+1}$.

Lemma 9.7. *Algorithm 1 will not add the children of node \mathbf{C} to our search queue if the maximum side length of \mathbf{C} is less than $\frac{\varepsilon D(\text{nn}_q, q)}{2\mu(\mu+1)\sqrt{d}}$.*

Proof. Let $\Delta(\mathbf{C})$ represent the maximum distance between any two points of cell \mathbf{C} .

By construction, we can expand \mathbf{C} only if some subcell of \mathbf{C} contains a point p such that $D(p, q) \leq (1 - \frac{\varepsilon}{2})D_{\text{near}}$. Note that since \mathbf{C} is examined, we have $D_{\text{near}} \leq D(\text{rep}_\mathbf{C}, q)$. Now assuming we expand \mathbf{C} , then we must have:

$$\begin{aligned} D(\text{rep}_\mathbf{C}, q) - D(p, q) &< \mu\Delta(\mathbf{C}) \\ D_{\text{near}} - (1 - \frac{\varepsilon}{2})D_{\text{near}} &< \mu\Delta(\mathbf{C}) \\ \frac{\varepsilon}{2}D_{\text{near}} &< \mu\Delta(\mathbf{C}) \\ \frac{\varepsilon}{2\mu}D_{\text{near}} &< \Delta(\mathbf{C}) \end{aligned}$$

Note that we substitute $D(\text{rep}_\mathbf{C}, q) < D_{\text{near}}$ and that by the definition of D_{near} as our candidate nearest neighbor distance, $D(\text{nn}_q, q) < D_{\text{near}}$. Our main modification from the symmetric case is that here $\Delta(\mathbf{C}) < (\mu+1)\sqrt{d}s$, where s is the maximum side length of \mathbf{C} , as opposed to $\sqrt{d}s$. Since cell \mathbf{C} may be covered by a left-ball of radius $\sqrt{d}s$ placed at a suitably chosen corner (as explained in Lemma 5.4), lemma 9.1 gives the required bound on $\Delta(\mathbf{C})$ \square

The main difference between this lemma and Lemma 7.4 is the extra factor of $\mu+1$ that we incur (as discussed) because of asymmetry. We only need do a little more work to obtain our final bounds:

Lemma 9.8. Given a cube G of side length D_{rough} , and letting $x = \frac{1}{\varepsilon^d} 2^d \mu^d (\mu + 1)^d d^{\frac{d}{2}} \left(\frac{D_{\text{rough}}}{D(\text{nn}_q, q)} \right)^d$ we can compute a $(1 + \varepsilon)$ -right sided nearest neighbor to q in $O(x^2 \log^d n)$ time.

Proof. Consider a quadtree search from q on a cube G of side length D_{rough} . By lemma 9.7, our algorithm will not expand cells with all side lengths smaller than $\varepsilon D(\text{nn}_q, q) / 2\mu(\mu + 1)\sqrt{d}$. But since the side length reduces by at least a factor of $\sqrt{2}$ in each dimension upon each split, all side lengths are less than this value after $k = \log_{\sqrt{2}} \left(2D_{\text{rough}}\mu(\mu + 1)\sqrt{d} / \varepsilon D(\text{nn}_q, q) \right)$ repeated bisections of our root cube. Observe now that $O(\log^d n)$ time is spent at each node by Lemma 9.6, that at the k -th level the number of nodes expanded is 2^{kd} , and that $\log_{\sqrt{2}} n = (\log_2 n)^2$. We then get a final time complexity bound of $O\left((1/\varepsilon^{2d}) 2^{2d} \mu^{2d} (\mu + 1)^{2d} d^d (D_{\text{rough}}/D(\text{nn}_q, q))^{2d} \log^d n\right)$. \square

Substituting $D_{\text{rough}} = \mu^2 \log(n) D(\text{nn}_q, q)$ in Lemma 9.8 gives us a bound of $O\left(2^{2d} \frac{1}{\varepsilon^{2d}} \mu^{6d} (\mu + 1)^{2d} d^d \log^{3d} n\right)$. This time is per cube of F that covers right-ball $B(q, D_{\text{rough}})$. Noting that there are 2^d such cubes gives us a final time complexity of $O\left(2^{3d} \frac{1}{\varepsilon^{2d}} \mu^{6d} (\mu + 1)^{2d} d^d \log^{3d} n\right)$. The space bound follows as in Section 7.

Logarithmic bounds for Asymmetric Bregman divergences We now extend our logarithmic bounds from Section 8 to asymmetric Bregman divergence $\sqrt{D_\phi}$. First note that the following Lemma goes through by identical argument to Lemma 8.1.

Lemma 9.9. Suppose we are given an interval $I = [x_1 x_2] \subset \mathbb{R}$ s.t. $x_1 < x_2$, $D_e(x_1, x_2) = r_e$ and $\sqrt{D_\phi(x_1, x_2)} = r_\phi$. Suppose we divide I into m subintervals of equal length with endpoints $x_1 = a_0 < a_1 < \dots < a_{m-1} < a_m = x_2$ where $D_e(a_i, a_{i+1}) = r_e/m$, for all $i \in [0..m-1]$. Then $\frac{r_\phi}{c_0 m} \leq \sqrt{D_\phi(a_i, a_{i+1})} \leq \frac{c_0 r_\phi}{m}$.

Corollary 9.1. If we recursively bisect an interval $I = [x_1 x_2] \subset \mathbb{R}$ s.t. $D_e(x_1, x_2) = r_e$ and $\sqrt{D_\phi(x_1, x_2)} = r_\phi$ into 2^i equal subintervals (under D_e), then $\frac{r_\phi}{c_0 2^i} \leq \sqrt{D_\phi(a_k, a_{k+1})} \leq \frac{c_0 r_\phi}{2^i}$ for any of the subintervals $[a_k a_{k+1}]$ so obtained. Hence after $i = \lceil \log \frac{c_0 r_\phi}{x} \rceil$ subdivisions, all intervals will be of length at most x .

We now construct a compressed Euclidean quad tree as before, modifying the Section 8 analysis slightly to account for the weaker packing bounds for $\sqrt{D_\phi}$ and the extra $\mu + 1$ factor on the diameter of a cell.

Theorem 9.1. Given an asymmetric decomposable Bregman divergence D_ϕ that is μ -defective over a domain with associated c_0 as in Section 8, we can compute a $(1 + \varepsilon)$ -approximate right-near-neighbor in time $O\left((\mu + 1)^d d^{\frac{d}{2}} \log n + \left(\frac{2c_0^4(\mu+1)\mu^3\sqrt{d}}{\varepsilon}\right)^d\right)$.

We note our first new Lemma, a slightly modified packing bound due to $\sqrt{D_\phi}$ not having a direct RTI.

Lemma 9.10. Given an interval $[x_1 x_2] \subset \mathbb{R}$ s.t. $\sqrt{D_\phi(x_1, x_2)} = r > 0$, and intervals with endpoints $a_0 < a_1 < \dots < a_{m-1} < a_m$, s.t. for all $i \in [0..m-1]$, $\sqrt{D_\phi(a_i, a_{i+1})} \geq l$, at most $O(\frac{c_0 r}{l})$ such intervals intersect $[x_1 x_2]$.

Proof. By the Lagrange form,

$$\frac{l}{r} < \frac{\sqrt{D_\phi(a_i, a_{i+1})}}{\sqrt{D_\phi(x_1, x_2)}} < c_0 \frac{D_e(a_i, a_{i+1})}{D_e(x_1, x_2)}, \quad (9.1)$$

or we can say that $\frac{D_e(a_i, a_{i+1})}{D_e(x_1, x_2)} > \frac{l}{rc_0}$. The RTI for D_e then gives us the required result. \square

Corollary 9.2. Given a ball B of radius r under $\sqrt{D_\phi}$, there can be at most $c_0^d (\frac{r}{l})^d$ disjoint cubes that can intersect B where each cube has side length at least l under $\sqrt{D_\phi}$.

As before, we find the smallest enclosing Bregman cube of side length s that encloses our point set, and then construct a compressed Euclidean quad-tree in pre-processing. Let L_i denote the cells at the i -th level.

Lemma 9.11. *Given a ball B of radius r under $\sqrt{D_\phi}$, let $i = \log \frac{s}{c_0 r}$. Then $|L_i \cap B| \leq O(c_0^d)$ and the side length of each cell in L_i is between r and $c_0^2 r$ under $\sqrt{D_{s\phi}}$. We can also explicitly retrieve the quadtree cells corresponding to $|L_i \cap B|$ in $O(c_0^d \log n)$ time.*

Proof. Note that for cells in L_i , we have side lengths between $\frac{s}{c_0 2^i}$ and $\frac{c_0 s}{2^i}$ by Corollary 9.1. Substituting $i = \log \frac{s}{c_0 r}$, these cells have side length between r and $c_0^2 r$ under $\sqrt{D_{s\phi}}$. Now, we look in each dimension at the number of disjoint intervals of length at least r that can intersect B . By Lemma 9.10, this is at most c_0 . The rest of the proof follows as in Lemma 8.2. \square

We first obtain in $O(\log n)$ time with our asymmetric ring-tree an $O(n)$ ANN q_{rough} to query point q , such that $\sqrt{D_\phi(q_{\text{rough}}, q)} = O(\mu^2 n \sqrt{D_\phi(\text{nn}_q, q)})$. We then use Lemma 9.11 to get $O(c_0^d)$ cells of our quadtree that intersect right ball $B(q, \sqrt{D_\phi(q_{\text{rough}}, q)})$.

Let us call this collection of cells as Q . We then carry out a quadtree search on each element of Q . Note that we expand only cells which may contain a point nearer to query point q than the current best candidate. We bound the depth of our search using μ -defectiveness similar to Lemma 8.4.

Lemma 9.12. *We need only expand cells of diameter greater than $\frac{\varepsilon \sqrt{D_\phi(\text{nn}_q, q)}}{2\mu}$*

Proof. By μ -defectiveness, similar to Lemma 7.4. \square

Corollary 9.3. *We will not expand cells where the length of each side is less than $x = \frac{\varepsilon \sqrt{D_\phi(\text{nn}_q, q)}}{2\mu(\mu+1)\sqrt{d}}$*

Proof. Note that a quadtree cell \mathbf{C} whose side length is less than x can be covered by a ball of radius $\sqrt{d}x$ under $\sqrt{D_\phi}$ with appropriately chosen corner as center of ball, as explained in proof of Lemma 5.4. Now by Lemma 9.1, $\sqrt{D_\phi(a, b)} \leq (\mu+1)\sqrt{d}x$, $\forall a, b \in \mathbf{C}$. Substituting for x from Lemma 9.12, the diameter of \mathbf{C} is at most $\frac{\varepsilon \sqrt{D_\phi(\text{nn}_q, q)}}{2\mu}$. \square

Let the spread be $\beta = \frac{D_{\text{rough}}}{\sqrt{D_{s\phi}(\text{nn}_q, q)}} = O(\mu^2 n)$.

Lemma 9.13. *We will only expand our tree to a depth of $k = \log(2c_0^3 \mu(\mu+1)\beta \sqrt{d}/\varepsilon)$.*

Lemma 9.14. *The number of cells expanded at the i -th level is $n_i < 2^d (\mu^d d^{\frac{d}{2}} c_0^{5d} + (\frac{c_0^{2i}}{\beta})^d)$.*

Proof. Recalling that the cells of Q start with side length at most $c_0^2 D_{\text{rough}}$, at the i -th level the side length of a cell \mathbf{C} is at most $\frac{c_0^3 D_{\text{rough}}}{2^i}$ under $\sqrt{D_\phi}$ by Corollary 9.1. And using Lemma 9.1, $\Delta \mathbf{C} < \sqrt{d}(\mu+1) \frac{c_0^3 D_{\text{rough}}}{2^i}$. Hence by μ -defectiveness there must be a point at distance at most $D_{\text{best}} = \sqrt{D_\phi(\text{nn}_q, q)} + \frac{\mu(\mu+1)c_0^3 \sqrt{d} D_{\text{rough}}}{2^i}$.

The side length of a cell \mathbf{C} at this level is at least $\frac{D_{\text{rough}}}{c_0 2^i}$, so the number of cells expanded is at most $n_i = c_0^d (\frac{D_{\text{best}}}{\Delta \mathbf{C}})^d = c_0^d (\mu(\mu+1)\sqrt{d}c_0^4 + \frac{c_0^{2i}}{\beta})^d$, by Corollary 9.2. Using the fact that $(a+b)^d < 2^d(a^d + b^d)$, we get $n_i < 2^d (\mu^d (\mu+1)^d d^{\frac{d}{2}} c_0^{5d} + (\frac{c_0^{2i}}{\beta})^d)$. \square

Simply summing up all i , the total number of nodes explored is

$$O(2^d \mu^d (\mu+1)^d c_0^{5d} \log(2c_0^3 \mu \beta \sqrt{d}/\varepsilon) + 2^{2d} c_0^{5d} \mu^d (\mu+1)^d d^{\frac{d}{2}} / \varepsilon^d)$$

, or

$$O\left(2^d \mu^d (\mu + 1)^d c_0^{5d} \log n + 2^{2d} c_0^{5d} \mu^d (\mu + 1)^d d^{\frac{d}{2}} / \epsilon^d\right)$$

, after substituting back for β and ignoring smaller terms. Recalling that there are c_0^d cells in Q adds a further c_0^d multiplicative factor.

10 Further work

A major open question is whether bounds independent of μ -defectiveness can be obtained for the complexity of ANN-search under Bregman divergences. As we have seen, traditional grid based methods rely heavily on the triangle inequality and packing bounds, and there are technical difficulties in adapting other method such as cone decompositions [12] or approximate Voronoi diagrams [21]. We expect that we will need to exploit geometry of Bregman divergences more substantially.

11 Acknowledgements

We thank Sarel Har-Peled and anonymous reviewers for helpful comments.

References

- [1] ACKERMANN, M., AND BLÖMER, J. Bregman clustering for separable instances. In *Algorithm Theory - SWAT 2010*, H. Kaplan, Ed., vol. 6139 of *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg, 2010, pp. 212–223.
- [2] ACKERMANN, M., BLÖMER, J., AND SCHOLZ, C. Hardness and non-approximability of bregman clustering problems. Tech. rep., Electronic colloquium on computational complexity, 2011. <http://eccc.hpi-web.de/report/2011/015/>.
- [3] ACKERMANN, M. R., AND BLÖMER, J. Coresets and approximate clustering for bregman divergences. In *Proceedings of the twentieth Annual ACM-SIAM Symposium on Discrete Algorithms* (Philadelphia, PA, USA, 2009), SODA '09, Society for Industrial and Applied Mathematics, pp. 1088–1097.
- [4] AFSHANI, P., ARGE, L., AND LARSEN, K. D. Orthogonal range reporting in three and higher dimensions. *Foundations of Computer Science, Annual IEEE Symposium on 0* (2009), 149–158.
- [5] AMARI, S., AND NAGAOKA, H. *Methods of Information Geometry*, vol. 191 of *Translations of Mathematical monographs*. Oxford University Press, 2000.
- [6] BANERJEE, A., MERUGU, S., DHILLON, I. S., AND GHOSH, J. Clustering with bregman divergences. *J. Mach. Learn. Res.* 6 (December 2005), 1705–1749.
- [7] BEYGELZIMER, A., KAKADE, S., AND LANGFORD, J. Cover trees for nearest neighbor. In *Proceedings of the 23rd international conference on Machine learning* (New York, NY, USA, 2006), ICML '06, ACM, pp. 97–104.
- [8] BOISSONNAT, J.-D., NIELSEN, F., AND NOCK, R. Bregman voronoi diagrams. *Discrete and Computational Geometry* 44 (2010), 281–307. 10.1007/s00454-010-9256-1.
- [9] BREGMAN, L. M. The Relaxation Method of Finding the Common Point of Convex Sets and Its Application to the Solution of Problems in Convex Programming. *USSR Computational Mathematics and Mathematical Physics* 7 (1967), 200–217.
- [10] CAYTON, L. Fast nearest neighbor retrieval for bregman divergences. In *Proceedings of the 25th international conference on Machine learning* (New York, NY, USA, 2008), ICML '08, ACM, pp. 112–119.
- [11] CAYTON, L. *Bregman proximity search*. PhD thesis, University of California, San Diego, 2009.
- [12] CHAN, T. M. Approximate nearest neighbor queries revisited. *Discrete and Computational Geometry* 20 (1998), 359–373. 10.1007/PL00009390.
- [13] CHAUDHURI, K., AND MCGREGOR, A. Finding metric structure in information theoretic clustering. In *COLT '08* (2008).
- [14] COLE, R., AND GOTTLIEB, L.-A. Searching dynamic point sets in spaces with bounded doubling dimension. In *Proceedings of the thirty-eighth annual ACM symposium on Theory of computing* (New York, NY, USA, 2006), STOC '06, ACM, pp. 574–583.
- [15] COLLINS, M., SCHAPIRE, R., AND SINGER, Y. Logistic regression, adaboost and bregman distances. *Machine Learning* 48, 1 (2002), 253–285.

- [16] COVER, T. M., AND THOMAS, J. A. *Elements of information theory*. Wiley-Interscience, New York, NY, USA, 1991.
- [17] CSISZÁR, I. I-Divergence Geometry of Probability Distributions and Minimization Problems. *The Annals of Probability* 3 (1975), 146–158.
- [18] EPPSTEIN, D., GOODRICH, M. T., AND SUN, J. Z. The skip quadtree: a simple dynamic data structure for multidimensional data. In *Proceedings of the twenty-first annual symposium on Computational geometry* (New York, NY, USA, 2005), SCG '05, ACM, pp. 296–305.
- [19] FARAGO, A., LINDER, T., AND LUGOSI, G. Fast nearest-neighbor search in dissimilarity spaces. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 15, 9 (sep 1993), 957–962.
- [20] GRAY, R., BUZO, A., GRAY JR, A., AND MATSUYAMA, Y. Distortion measures for speech processing. *Acoustics, Speech and Signal Processing, IEEE Transactions on* 28, 4 (Aug 1980), 367–376.
- [21] HAR-PELED, S. A replacement for voronoi diagrams of near linear size. In *Proceedings of the 42nd IEEE symposium on Foundations of Computer Science* (Washington, DC, USA, 2001), FOCS '01, IEEE Computer Society, pp. 94–105.
- [22] HAR-PELED, S., AND MAZUMDAR, S. Fast algorithms for computing the smallest k-enclosing circle. *Algorithmica* 41 (2005), 147–157. 10.1007/s00453-004-1123-0.
- [23] HAR-PELED, S., AND MENDEL, M. Fast construction of nets in low dimensional metrics, and their applications. In *Proceedings of the twenty-first annual symposium on Computational geometry* (New York, NY, USA, 2005), SCG '05, ACM, pp. 150–158.
- [24] HJALTASON, G. R., AND SAMET, H. Index-driven similarity search in metric spaces (survey article). *ACM Trans. Database Syst.* 28 (December 2003), 517–580.
- [25] INDYK, P., AND MOTWANI, R. Approximate nearest neighbors: towards removing the curse of dimensionality. In *Proceedings of the thirtieth annual ACM symposium on Theory of computing* (New York, NY, USA, 1998), STOC '98, ACM, pp. 604–613.
- [26] KRAUTHGAMER, R., AND LEE, J. The black-box complexity of nearest neighbor search. In *Automata, Languages and Programming*, J. Daz, J. Karhumki, A. Lepist, and D. Sannella, Eds., vol. 3142 of *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg, 2004, pp. 153–178.
- [27] MAHALANOBIS, P. C. On the generalised distance in statistics. *Proc. National Institute of Sciences in India* 2, 1 (1936), 49–55.
- [28] MANTHEY, B., AND RÖGLIN, H. Worst-case and smoothed analysis of k-means clustering with bregman divergences. In *Algorithms and Computation*, Y. Dong, D.-Z. Du, and O. Ibarra, Eds., vol. 5878 of *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg, 2009, pp. 1024–1033.
- [29] MORALEDA, J., SHAKHNAROVICH, DARRELL, T., AND INDYK, P. Nearest-neighbors methods in learning and vision. theory and practice. *Pattern Analysis and Applications* 11 (2008), 221–222. 10.1007/s10044-007-0076-8.
- [30] NIELSEN, F., AND BOLTZ, S. The burbea-rao and bhattacharyya centroids. *CoRR abs/1004.5049* (2010).
- [31] NIELSEN, F., AND NOCK, R. On the smallest enclosing information disk. *Information Processing Letters* 105, 3 (2008), 93 – 97.

- [32] NIELSEN, F., AND NOCK, R. Sided and symmetrized bregman centroids. *IEEE Trans. Inf. Theor.* 55 (June 2009), 2882–2904.
- [33] NIELSEN, F., PIRO, P., AND BARLAUD, M. Bregman vantage point trees for efficient nearest neighbor queries. In *Proceedings of the 2009 IEEE international conference on Multimedia and Expo* (Piscataway, NJ, USA, 2009), ICME'09, IEEE Press, pp. 878–881.
- [34] NIELSEN, F., PIRO, P., AND BARLAUD, M. Tailored bregman ball trees for effective nearest neighbors. In *European Workshop on Computational Geometry* (2009).
- [35] SARIEL-HAR-PELED. *Geometric Approximation Algorithms*. AMS, 2011. <http://goo.gl/pLiEO>.
- [36] SPELLMAN, E., AND VEMURI, B. Efficient shape indexing using an information theoretic representation. In *Image and Video Retrieval*, W.-K. Leow, M. Lew, T.-S. Chua, W.-Y. Ma, L. Chaisorn, and E. Bakker, Eds., vol. 3568 of *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg, 2005, pp. 590–590.
- [37] ZHANG, Z., OOI, B. C., PARTHASARATHY, S., AND TUNG, A. K. H. Similarity search on bregman divergence: towards non-metric indexing. *Proc. VLDB Endow.* 2 (August 2009), 13–24.

A Proofs from Section 4

We present here full versions of proofs of μ -defectiveness and RTI from Section 4.

A.1 Proof of RTI for $\sqrt{D_{s\phi}}$

Lemma A.1. $\sqrt{D_{s\phi}}$ satisfies the reverse triangle inequality.

Proof. Fix $a \leq x \leq b$, and assume that the reverse triangle inequality does not hold:

$$\begin{aligned} \sqrt{D_{s\phi}(a, x)} + \sqrt{D_{s\phi}(x, b)} &> \sqrt{D_{s\phi}(a, b)} \\ \sqrt{(x-a)(\phi'(x) - \phi'(a))} + \sqrt{(b-x)(\phi'(b) - \phi'(x))} &> \sqrt{(b-a)(\phi'(b) - \phi'(a))} \end{aligned}$$

Squaring both sides, we get:

$$\begin{aligned} (x-a)(\phi'(x) - \phi'(a)) + (b-x)(\phi'(b) - \phi'(x)) \\ + 2\sqrt{(x-a)(b-x)(\phi'(x) - \phi'(a))(\phi'(b) - \phi'(x))} &> (b-a)(\phi'(b) - \phi'(a)) \\ (b-x)(\phi'(x) - \phi'(a)) + (x-a)(\phi'(b) - \phi'(x)) \\ - 2\sqrt{(x-a)(b-x)(\phi'(x) - \phi'(a))(\phi'(b) - \phi'(x))} &< 0 \\ \left(\sqrt{(b-x)(\phi'(x) - \phi'(a))} - \sqrt{(x-a)(\phi'(b) - \phi'(x))} \right)^2 &< 0 \end{aligned}$$

which is a contradiction, since the LHS is a perfect square. □

A.2 Proof of Lemma 4.3

Lemma A.2. Given any interval $I = [x_1, x_2]$ on the real line, there exists a finite μ such that $\sqrt{D_{s\phi}}$ is μ -defective with respect to I .

Proof. Consider three points $a, b, q \in I$.

Due to symmetry of the cases and conditions, there are three cases to consider: $a < q < b$, $a < b < q$ and $q < b < a$.

Case 1: Here $a < q < b$. The following is trivially true by the monotonicity of $\sqrt{D_{s\phi}}$.

$$\left| \sqrt{D_{s\phi}(q, a)} - \sqrt{D_{s\phi}(q, b)} \right| < \sqrt{D_{s\phi}(a, b)} \quad (\text{A.1})$$

Cases 2 and 3: For the remaining symmetric cases, $a < b < q$ and $q < b < a$, note that since $\sqrt{D_{s\phi}(q, a)} - \sqrt{D_{s\phi}(q, b)}$ and $\sqrt{D_{s\phi}(a, b)}$ are both bounded, continuous functions on a compact domain (the interval $[x_1, x_2]$), we need only show that the following limit exists:

$$\lim_{a \rightarrow b} \frac{|\sqrt{D_{s\phi}(q, a)} - \sqrt{D_{s\phi}(q, b)}|}{\sqrt{D_{s\phi}(a, b)}} \quad (\text{A.2})$$

First consider $a < b < q$, and we assume $\lim_{b \rightarrow a}$. We will use the following substitutions repeatedly in our derivation: $b = a + h$, $\lim_{h \rightarrow 0} \phi(a + h) = \lim_{h \rightarrow 0} (\phi(a) + h\phi'(a))$, and $\lim_{h \rightarrow 0} \sqrt{1 + h} = \lim_{h \rightarrow 0} (1 + h/2)$. For ease of computation, we replace ϕ' by ψ , to be restored at the last step.

$$\lim_{a \rightarrow b} \frac{\sqrt{D_{s\phi}(a, q)} - \sqrt{D_{s\phi}(b, q)}}{\sqrt{D_{s\phi}(a, b)}} = \frac{\lim_{a \rightarrow b} (\sqrt{(q-a)(\psi(q) - \psi(a))} - \sqrt{(q-b)(\psi(q) - \psi(b))})}{\lim_{a \rightarrow b} \sqrt{(b-a)(\psi(b) - \psi(a))}} \quad (\text{A.3})$$

Computing the denominator:

$$\begin{aligned} \lim_{b \rightarrow a} \sqrt{(b-a)(\psi(b) - \psi(a))} &= \lim_{h \rightarrow 0} \sqrt{(a+h-a)(\psi(a+h) - \psi(a))} \\ &= \lim_{h \rightarrow 0} \sqrt{h(\psi(a) + h\psi'(a) - \psi(a))} \\ &= \lim_{h \rightarrow 0} \sqrt{h(h\psi'(a))} = \lim_{h \rightarrow 0} h\sqrt{\psi'(a)} \end{aligned}$$

We now address the numerator:

$$\begin{aligned} \lim_{b \rightarrow a} \sqrt{(q-a)(\psi(q) - \psi(a))} - \sqrt{(q-b)(\psi(q) - \psi(b))} \\ &= \lim_{h \rightarrow 0} \sqrt{(q-a)(\psi(q) - \psi(a))} - \sqrt{(q-a-h)(\psi(q) - \psi(a) - h\psi'(a))} \\ &= \lim_{h \rightarrow 0} \sqrt{(q-a)(\psi(q) - \psi(a))} - \sqrt{(q-a) \left(1 - \frac{h}{q-a}\right) (\psi(q) - \psi(a)) \left(1 - h \frac{\psi'(a)}{\psi(q) - \psi(a)}\right)} \\ &= \lim_{h \rightarrow 0} \sqrt{(\psi(q) - \psi(a))(q-a)} \left(1 - \sqrt{1 - \frac{h}{q-a}} \sqrt{1 - h \frac{\psi'(a)}{\psi(q) - \psi(a)}}\right) \\ &= \lim_{h \rightarrow 0} \sqrt{(\psi(q) - \psi(a))(q-a)} \left(1 - \left(1 - \frac{h}{2(q-a)}\right) \left(1 - h \frac{\psi'(a)}{2(\psi(q) - \psi(a))}\right)\right) \\ &= \lim_{h \rightarrow 0} \sqrt{(\psi(q) - \psi(a))(q-a)} \left(\frac{h}{2(q-a)} + h \frac{\psi'(a)}{2(\psi(q) - \psi(a))} - \frac{h^2}{4(q-a)(\psi(q) - \psi(a))}\right) \end{aligned}$$

Dropping higher order terms of h , the above reduces to:

$$\lim_{h \rightarrow 0} h \sqrt{(\psi(q) - \psi(a))(q-a)} \left(\frac{1}{2(q-a)} + \frac{\psi'(a)}{2(\psi(q) - \psi(a))}\right)$$

Now combine numerator and denominator back in equation A.3.

$$\begin{aligned}
\lim_{b \rightarrow a} \frac{\sqrt{D_{s\phi}(a,q)} - \sqrt{D_{s\phi}(b,q)}}{\sqrt{D_{s\phi}(a,b)}} &= \frac{\lim_{h \rightarrow 0} h \sqrt{(\psi(q) - \psi(a))(q-a)} \left(\frac{1}{2(q-a)} + \frac{\psi'(a)}{2(\psi(q) - \psi(a))} \right)}{\lim_{h \rightarrow 0} h \sqrt{\psi'(a)}} \\
&= \sqrt{\frac{(\psi(q) - \psi(a))(q-a)}{\psi'(a)}} \left(\frac{1}{2(q-a)} + \frac{\psi'(a)}{2(\psi(q) - \psi(a))} \right) \\
&= \frac{1}{2} \left(\sqrt{\frac{\psi(q) - \psi(a)}{\psi'(a)(q-a)}} + \sqrt{\frac{\psi'(a)(q-a)}{\psi(q) - \psi(a)}} \right)
\end{aligned}$$

Substituting back $\phi'(x)$ for $\psi(x)$, we see that limit A.2 exists, provided ϕ is strictly convex:

$$\frac{1}{2} \left(\sqrt{\frac{\phi'(q) - \phi'(a)}{\phi''(a)(q-a)}} + \sqrt{\frac{\phi''(a)(q-a)}{\phi'(q) - \phi'(a)}} \right) \quad (\text{A.4})$$

The analysis follows symmetrically for case 3, where $q < b < a$. □

We now show that right-sided μ -defectiveness holds for D_ϕ . To show this, we need to establish the following relationship between $D_\phi(a, b)$ and $D_\phi(b, a)$ over a bounded interval $I \subset \mathbb{R}$.

A.3 Proof of Lemma 4.4

Lemma A.3. *Given a Bregman divergence D_ϕ and a bounded interval $I \subset \mathbb{R}$, we have that $\sqrt{D_\phi(a, b)}/\sqrt{D_\phi(b, a)}$ is bounded by some constant $c_0 \forall a, b \in I$ where c_0 depends on the choice of divergence and interval.*

Proof. By continuity and compactness, over a finite interval I we have that $c_0 = \max_x \phi_i''(x)/\min_y \phi_i''(y)$ is bounded. Now by using the Lagrange form of $\sqrt{D_\phi(a, b)}$, we get that $\sqrt{D_\phi(a, b)}/\sqrt{D_\phi(b, a)} < \sqrt{c_0}$ □

A.4 Proof that $\sqrt{D_\phi}$ is μ -defective

Lemma A.4. *Given any interval $I = [x_1, x_2]$ on the real line, there exists a finite μ such that $\sqrt{D_\phi}$ is right-sided μ -defective with respect to I*

Proof. Consider any three points $a, b, q \in I$. We will prove that there exists finite μ such that:

$$\left| \sqrt{D_\phi(a, q)} - \sqrt{D_\phi(b, q)} \right| < \mu \sqrt{D_\phi(b, a)} \quad (\text{A.5})$$

Here there are now six cases to consider: $a < q < b$, $b < q < a$, $a < b < q$, $b < a < q$, $q < b < a$, and $q < a < b$.

Case 1 and 2: Here $a < q < b$. By monotonicity we have that:

$$\left| \sqrt{D_\phi(a, q)} - \sqrt{D_\phi(b, q)} \right| < \sqrt{D_\phi(a, b)} + \sqrt{D_\phi(b, a)} \quad (\text{A.6})$$

But by lemma 4.4, we have that $\sqrt{D_\phi(a, b)} < c \sqrt{D_\phi(b, a)}$ for some constant c defined over I . This implies that $\left| \sqrt{D_\phi(a, q)} - \sqrt{D_\phi(b, q)} \right| / \sqrt{D_\phi(b, a)} < c + 1$, i.e, it is bounded over I . A similar analysis works for Case 2 where $b < q < a$.

Cases 3 and 4: For these two cases, $a < b < q$ and $b < a < q$, note that since $\sqrt{D_\phi(q, a)} - \sqrt{D_\phi(q, b)}$ and $\sqrt{D_\phi(b, a)}$ are both bounded, continuous functions on a compact domain (the interval $[x_1, x_2]$), we need only show that the following limit exists:

$$\lim_{a \rightarrow b} \frac{|\sqrt{D_\phi(a, q)} - \sqrt{D_\phi(b, q)}|}{\sqrt{D_\phi(b, a)}} \quad (\text{A.7})$$

First consider $a < b < q$, and we assume $\lim_{b \rightarrow a}$. We will use the following substitutions repeatedly in our derivation: $b = a + h$, $\lim_{h \rightarrow 0} \phi(a + h) = \lim_{h \rightarrow 0} (\phi(a) + h\phi'(a))$, $\lim_{h \rightarrow 0} \phi(b) = \phi(a + h) = \lim_{h \rightarrow 0} (\phi(a) + h\psi(a) + \frac{h^2\psi'(a)}{2})$ and $\lim_{h \rightarrow 0} \sqrt{1+h} = \lim_{h \rightarrow 0} (1 + h/2)$. For ease of computation, we replace ϕ' by ψ , to be restored at the last step.

$$\lim_{a \rightarrow b} \frac{\sqrt{D_\phi(a, q)} - \sqrt{D_\phi(b, q)}}{\sqrt{D_\phi(b, a)}} = \lim_{a \rightarrow b} \frac{\sqrt{\phi(a) - \phi(q) - \psi(q)(a - q)} - \sqrt{\phi(b) - \phi(q) - \psi(q)(b - q)}}{\sqrt{\phi(b) - \phi(a) - \psi(a)(b - a)}} \quad (\text{A.8})$$

Computing the denominator:

$$\begin{aligned} \lim_{a \rightarrow b} \sqrt{\phi(b) - \phi(a) - \psi(a)(b - a)} &= \lim_{h \rightarrow 0} \sqrt{\phi(a) + h\psi(a) + \frac{h^2\psi'(a)}{2} - \phi(a) - h\psi(a)} \\ &= \lim_{h \rightarrow 0} \sqrt{\frac{h^2\psi'(a)}{2}} \\ &= \lim_{h \rightarrow 0} h \sqrt{\frac{\psi'(a)}{2}} \end{aligned}$$

We now address the numerator:

$$\begin{aligned} &\lim_{a \rightarrow b} \left(\sqrt{\phi(a) - \phi(q) - \psi(q)(a - q)} - \sqrt{\phi(b) - \phi(q) - \psi(q)(b - q)} \right) \\ &= \lim_{h \rightarrow 0} \sqrt{\phi(a) - \phi(q) - \psi(q)(a - q)} - \sqrt{\phi(a) - \phi(q) - \psi(q)(a - q) + h(\psi(a) - \psi(q))} \\ &= \lim_{h \rightarrow 0} \sqrt{D_\phi(a, q)} - \sqrt{D_\phi(a, q) \left(1 + \frac{h(\psi(a) - \psi(q))}{D_\phi(a, q)} \right)} \\ &= \lim_{h \rightarrow 0} \sqrt{D_\phi(a, q)} \left(1 - \sqrt{1 - \frac{h(\psi(q) - \psi(a))}{D_\phi(a, q)}} \right) \\ &= \lim_{h \rightarrow 0} \sqrt{D_\phi(a, q)} \left(1 - \left(1 - \frac{h(\psi(q) - \psi(a))}{2D_\phi(a, q)} \right) \right) \\ &= \lim_{h \rightarrow 0} \frac{h(\psi(q) - \psi(a))}{2\sqrt{D_\phi(a, q)}} \end{aligned}$$

Now combine numerator and denominator back in equation A.8, and note that $D_\phi(a, q) = \frac{1}{2}(\psi'(x))(q - a)^2$, for some $x \in [ab]$.

$$\begin{aligned} \lim_{a \rightarrow b} \frac{\sqrt{D_\phi(a, q)} - \sqrt{D_\phi(b, q)}}{\sqrt{D_\phi(b, a)}} &= \frac{\lim_{h \rightarrow 0} \frac{h(\psi(q) - \psi(a))}{2\sqrt{D_\phi(a, q)}}}{\lim_{h \rightarrow 0} h \sqrt{\frac{\psi'(a)}{2}}} \\ &= \frac{(\psi(q) - \psi(a))}{q - a} \frac{\sqrt{\psi'(a)}}{\sqrt{\psi'(x)}} \end{aligned}$$

Substituting back $\phi'(x)$ for $\psi(x)$, we see that limit A.7 exists, provided ϕ is strictly convex:

$$\frac{(\phi'(q) - \phi'(a))}{q - a} \frac{\sqrt{\phi''(a)}}{\sqrt{\phi''(x)}} \quad (\text{A.9})$$

The analysis follows symmetrically for case 4, by noting that $\lim_{a \rightarrow b} \frac{\sqrt{D_\phi(a,b)}}{\sqrt{D_\phi(b,a)}} = 1$ and that $\sqrt{D_\phi(a,q)} - \sqrt{D_\phi(b,q)} = -(\sqrt{D_\phi(b,q)} - \sqrt{D_\phi(a,q)})$, i.e we may suitably interchange a and b .

Cases 5 and 6: Here $q < a < b$ or $q < b < a$. Looking more carefully at the analysis for cases 3 and 4, note that the ordering $q < a < b$ vs $a < b < q$ does not affect the magnitude of the expression for limit A.7, only the sign. Hence we can use the same analysis to prove μ -defectiveness for cases 5 and 6. \square

Corollary A.1. *Given any interval $I = [x_1, x_2]$ on the real line, there exists a finite μ such that $\sqrt{D_\phi}$ is left-sided μ -defective with respect to I*

Proof. Follows from similar computation. \square

We extend our results to d dimensions naturally now by showing that if M is a domain such that $\sqrt{D_{s\phi}}$ and $\sqrt{D_\phi}$ are μ -defective with respect to the projection of M onto each coordinate axis, then $\sqrt{D_{s\phi}}$ and $\sqrt{D_\phi}$ are μ -defective with respect to all of M .

A.5 Proof of μ -defectiveness in d dimensions

Lemma A.5. *Consider three points, $a = (a_1, \dots, a_i, \dots, a_d)$, $b = (b_1, \dots, b_i, \dots, b_d)$, $q = (q_1, \dots, q_i, \dots, q_d)$ such that $|\sqrt{D_{s\phi}(a_i, q_i)} - \sqrt{D_{s\phi}(b_i, q_i)}| < \mu \sqrt{D_{s\phi}(a_i, b_i)}$, $\forall 1 \leq i \leq d$. Then*

$$\left| \sqrt{D_{s\phi}(a, q)} - \sqrt{D_{s\phi}(b, q)} \right| < \mu \sqrt{D_{s\phi}(a, b)} \quad (\text{A.10})$$

Similarly, if $|\sqrt{D_\phi(a_i, q_i)} - \sqrt{D_\phi(b_i, q_i)}| < \mu \sqrt{D_\phi(a_i, b_i)}$, $\forall 1 \leq i \leq d$. Then

$$\left| \sqrt{D_\phi(a, q)} - \sqrt{D_\phi(b, q)} \right| < \mu \sqrt{D_\phi(b, a)} \quad (\text{A.11})$$

Proof.

$$\begin{aligned} & \left| \sqrt{D_{s\phi}(a, q)} - \sqrt{D_{s\phi}(b, q)} \right| < \mu \sqrt{D_{s\phi}(a, b)} \\ & D_{s\phi}(a, q) + D_{s\phi}(b, q) - 2\sqrt{D_{s\phi}(a, q)D_{s\phi}(b, q)} < \mu^2 D_{s\phi}(a, b) \\ & \sum_{i=1}^d (D_{s\phi}(a_i, q_i) + D_{s\phi}(b_i, q_i)) - 2\sqrt{D_{s\phi}(a, q)D_{s\phi}(b, q)} < \mu^2 \sum_{i=1}^d D_{s\phi}(a_i, b_i) \\ & \sum_{i=1}^d (D_{s\phi}(a_i, q_i) + D_{s\phi}(b_i, q_i) - \mu^2 D_{s\phi}(a_i, b_i)) < 2\sqrt{D_{s\phi}(a, q)D_{s\phi}(b, q)} \end{aligned}$$

The last inequality is what we need to prove for μ -defectiveness with respect to a, b, q . By assumption we already have μ -defectiveness w.r.t each a_i, b_i, q_i , for every $1 \leq i \leq d$:

$$\begin{aligned} & D_{s\phi}(a_i, q_i) + D_{s\phi}(b_i, q_i) - \mu^2 D_{s\phi}(a_i, b_i) < 2\sqrt{D_{s\phi}(a_i, q_i)D_{s\phi}(b_i, q_i)} \\ & \sum_{i=1}^d (D_{s\phi}(a_i, q_i) + D_{s\phi}(b_i, q_i) - \mu^2 D_{s\phi}(a_i, b_i)) < 2\sum_{i=1}^d \sqrt{D_{s\phi}(a_i, q_i)D_{s\phi}(b_i, q_i)} \end{aligned}$$

So to complete our proof we need only show:

$$\sum_{i=1}^d \sqrt{D_{s\phi}(a_i, q_i)} \sqrt{D_{s\phi}(b_i, q_i)} \leq \sqrt{D_{s\phi}(a, q)} \sqrt{D_{s\phi}(b, q)} \quad (\text{A.12})$$

But notice the following:

$$\begin{aligned} \sqrt{D_{s\phi}(a, q)} &= \left(\sum_{i=1}^d D_{s\phi}(a_i, q_i) \right)^{\frac{1}{2}} = \left(\sum_{i=1}^d \left(\sqrt{D_{s\phi}(a_i, q_i)} \right)^2 \right)^{\frac{1}{2}} \\ \sqrt{D_{s\phi}(b, q)} &= \left(\sum_{i=1}^d D_{s\phi}(b_i, q_i) \right)^{\frac{1}{2}} = \left(\sum_{i=1}^d \left(\sqrt{D_{s\phi}(b_i, q_i)} \right)^2 \right)^{\frac{1}{2}} \end{aligned}$$

So inequality A.12 is simply a form of the Cauchy-Schwarz inequality, which states that for two vectors u and v in \mathbb{R}^d , that $|\langle u, v \rangle| \leq \|u\| \|v\|$, or that

$$\left| \sum_{i=1}^d u_i v_i \right| \leq \left(\sum_{i=1}^d u_i^2 \right)^{\frac{1}{2}} \left(\sum_{i=1}^d v_i^2 \right)^{\frac{1}{2}}$$

The second part of the proposition can be derived by an essentially identical argument. \square

B Numerical arguments for bisection

In our algorithms, we are required to *bisect* a given interval with respect to the distance measure D , as well as construct points that lie a fixed distance away from a given point. We note that in both these operations, we do not need exact answers: a constant factor approximation suffices to preserve all asymptotic bounds. In particular, our algorithms assume two procedures:

1. Given interval $[ab] \subset \mathbb{R}$, find $\bar{x} \in [ab]$ such that $(1 - \alpha) \sqrt{D_{s\phi}(a, \bar{x})} < \sqrt{D_{s\phi}(\bar{x}, b)} < (1 + \alpha) \sqrt{D_{s\phi}(a, \bar{x})}$ to a lesser degree by
2. Given $q \in \mathbb{R}$ and distance r , find \bar{x} s.t $|\sqrt{D_{s\phi}(q, \bar{x})} - r| < \alpha r$

For a given $\sqrt{D_{s\phi}} : \mathbb{R} \rightarrow \mathbb{R}$ and precision parameter $0 < \alpha < 1$, we describe a procedure that yields an $0 < \alpha < 1$ approximation in $O(\log c_0 + \log \mu + \log \frac{1}{\alpha})$ steps for both problems, where c_0 implicitly depends on the domain of convex function ϕ :

$$c_0 = \sqrt{\max_{1 \leq i \leq d} \left(\max_x \phi_i''(x) / \min_y \phi_i''(y) \right)} \quad (\text{B.1})$$

Note that this implies linear convergence. While more involved numerical methods such as Newton's method may yield better results, our approximation algorithm serve as proof-of-concept that the numerical precision is not problematic.

A careful adjustment of our NN-analysis now gives a $O\left((\log \mu + \log c_0 + \log \frac{1}{\alpha}) 2^{2d} (1 + \alpha)^d \frac{1}{\epsilon^d} \mu^{3d} d^{\frac{d}{2}} \log^{2d} n\right)$ time complexity to compute a $(1 + \epsilon)$ -ANN to query point q .

We now describe some useful properties of $D_{s\phi}$.

Lemma B.1. Consider $\sqrt{D_{s\phi}} : \mathbb{R} \rightarrow \mathbb{R}$ such that $c_0 = \sqrt{\max_x \phi''(x) / \min_y \phi''(y)}$. Then for any two intervals $[x_1 x_2], [x_3 x_4] \subset \mathbb{R}$,

$$\frac{1}{c_0} \frac{|x_1 - x_2|}{|x_3 - x_4|} < \frac{\sqrt{D_{s\phi}(x_1, x_2)}}{\sqrt{D_{s\phi}(x_3, x_4)}} < c_0 \frac{|x_1 - x_2|}{|x_3 - x_4|} \quad (\text{B.2})$$

Proof. The lemma follows by the definition of c_0 and by direct computation from the Lagrange form of $\sqrt{D_{s\phi}(a,b)}$, i.e., $\sqrt{D_{s\phi}(a,b)} = \sqrt{\phi''(\bar{x}_{ab})}|b-a|$, for some $\bar{x}_{ab} \in [ab]$. \square

Lemma B.2. *Given a point $q \in \mathbb{R}$, distance $r \in \mathbb{R}$, precision parameter $0 < \alpha < 1$ and a μ -defective $\sqrt{D_{s\phi}} : \mathbb{R} \rightarrow \mathbb{R}$, we can locate a point x_i such that $|\sqrt{D_{s\phi}(q, x_i)} - r| < \alpha r$ in $O(\log \frac{1}{\alpha} + \log \mu + \log c_0)$ time.*

Proof. Let x be the point such that $\sqrt{D_{s\phi}(q, x)} = r$. We outline an iterative process, 2, with i -th iterate x_i that converges to x . First note that $\frac{\sqrt{\phi''(q)}}{c_0} \leq \sqrt{\min_y \phi''(y)}$ and $\frac{\sqrt{\phi''(q)}}{c_0} \geq \frac{\max_z \sqrt{\phi''(z)}}{c_0^2}$. It immediately follows that

Algorithm 2 QueryApproxDist(q, r, c_0, α)

Let $x_0 > q$ be such that $\frac{\sqrt{\phi''(q)}}{c_0}(x_0 - q) = r$

Let step = $(x_0 - q)/2$

repeat

if $\sqrt{D_{s\phi}(q, x_i)} < r$ **then**

$x_{i+1} = x_i + \text{step}$

else

$x_{i+1} = x_i - \text{step}$

end if

 step = step/2

until $|\sqrt{D_{s\phi}(q, x_i)} - r| \leq \alpha r$

Return $\bar{x} = x_i$

$$r \leq \sqrt{D_{s\phi}(q, x_0)} \leq c_0^2 r.$$

By construction, $|x_i - x| \leq |x_0 - q|/2^i$. Hence by Lemma B.1, $\sqrt{D_{s\phi}(x_i, x)} < \frac{c_0^3 r}{2^i}$. We now use μ -defectiveness to upper bound our error $|\sqrt{D_{s\phi}(q, x_i)} - \sqrt{D_{s\phi}(q, x)}|$ at the i -th iteration:

$$\left| \sqrt{D_{s\phi}(q, x_i)} - \sqrt{D_{s\phi}(q, x)} \right| < \frac{\mu c_0^3 r}{2^i} \quad (\text{B.3})$$

Choosing i such that $(\mu c_0^3)/2^i \leq \alpha$ implies that $i \leq \log \frac{1}{\alpha} + \log \mu + 3 \log c_0$. \square

An almost identical procedure can locate an approximate bisection point of interval $[ab]$ in $O(\log \mu + \log c_0 + \log \frac{1}{\alpha})$ time, and similar techniques can be applied for $\sqrt{D_\phi}$. We omit the details here.